Adv. Appl. Probab. 1–31 (2024)
*doi:*10.1017/apr.2024.4

TOPOLOGICAL RECONSTRUCTION OF COMPACT SUPPORTS OF DEPENDENT STATIONARY RANDOM VARIABLES

SADOK KALLEL,* *American University of Sharjah*
SANA LOUHICHI,** *Université Grenoble Alpes*

Abstract

In this paper we extend results on reconstruction of probabilistic supports of independent and identically distributed random variables to supports of dependent stationary $\mathbb{R}^d$-valued random variables. All supports are assumed to be compact of positive reach in Euclidean space. Our main results involve the study of the convergence in the Hausdorff sense of a cloud of stationary dependent random vectors to their common support. A novel topological reconstruction result is stated, and a number of illustrative examples are presented. The example of the Möbius Markov chain on the circle is treated at the end with simulations.

Keywords: Hausdorff distance; stationary dependent random variables; mixing; Markov chains; compact support; concentration; positive reach; topological inference

2020 Mathematics Subject Classification: Primary 60F99; 60G10
Secondary 62G05; 55P10

1. Introduction

Given a sequence of stationary random variables of unknown common law and unknown compact support $\mathbb{M}$ (Section 3), it can be very useful in practice to identify topological properties of $\mathbb{M}$ based on observed data. Data analysis in high-dimensional spaces from a probabilistic point of view was initiated in [33], where data was assumed to be drawn from sampling an independent and identically distributed (i.i.d.) probability distribution on (or near) a submanifold $\mathbb{M}$ of Euclidean space. Topological properties of $\mathbb{M}$ (homotopy type and homology) were deduced based on the random samples and the geometrical properties of $\mathbb{M}$. Several papers on probability and topological inference ensued, some taking a persistence homology approach by providing a confidence set for persistence diagrams corresponding to the Hausdorff distance of a sample from a distribution supported on $\mathbb{M}$ [16].

Topology intervenes in probability through reconstruction results (see [3, 7, 8, 16, 29, 33] and references therein). This research direction is now recognized as part of 'manifold learning'. Given an n-point cloud $\mathbb{X}_n$ lying in a support $\mathbb{M}$, which is generally assumed to be a compact subspace of $\mathbb{R}^d$ for some $d > 0$, and given a certain probability distribution of these n

Received 6 April 2022; accepted 30 December 2023.
* Postal address: American University of Sharjah, UAE, and Laboratoire Painlevé, Université de Lille, France. Email address: skallel@aus.edu
** Postal address: Université Grenoble Alpes, CNRS, Grenoble INP, LJK 38000 Grenoble, France. Email address: sana.louhichi@univ-grenoble-alpes.fr

points on $\mathbb{M}$, one can formulate from this data practical conditions to reconstruct, up to homotopy or up to homology, the support $\mathbb{M}$. Reconstruction up to homotopy means recovering the homotopy type of $\mathbb{M}$. Reconstruction up to homology means determining, up to a certain degree, the homology groups of $\mathbb{M}$. Recovering the geometry of $\mathbb{M}$, including curvature and volume, is a much more delicate task (see [1, 13, 18, 33, 39]).

The goal of our work is to extend work of Nigoyi, Smale and Weinberger [33] from data drawn from sampling an i.i.d. probability distribution that has support on a smooth submanifold $\mathbb{M}$ of Euclidean space to data drawn from stationary *dependent* random variables concentrated inside a compact space of *positive reach* (or PR set). It is fitting here to define this notion: the *reach* of a closed set S in a metric space is the supremum $\tau \geq 0$ such that any point within distance less than τ of S has a unique nearest point in S. Spaces of positive reach τ were introduced in [17]; they form a natural family of spaces that are more general than convex sets ($\tau = \infty$) or smooth submanifolds, but share many of their common integro-geometric properties, such as 'curvature measures' [38] (see Section 2).

The interest in going beyond independence lies in the fact that many observations in everyday life are dependent, and independence is not sufficient to describe these phenomena. The study of the data support topologically and geometrically in this case can be instrumental in directional statistics, for example, where the observations are often correlated. This can provide information on animal migration paths or wind directions, for instance. Modeling by a Markov chain on an unknown compact manifold, with or without boundary, makes it possible to study such examples. Other illustrative examples can be found in more applied fields, for instance in cosmology [9], medicine [20], imaging [37], biology [2], and environmental science [22].

To get information on an unknown support from stationary dependent data, we need to study the convergence in the Hausdorff distance d_H of the data, seen as a (finite) point cloud, to its support, similarly to what was done in the i.i.d. case [7, 10, 11, 16]. The main feature of interest in the metric d_H is the following property: if $S \subset M$ is this point cloud in M, then $d_H(S, M) \leq \epsilon$ is equivalent to S being ϵ-dense in M (see Section 2). We can expand this relationship to the random case as follows.

Definition 1.1. We say that a point cloud $\mathbb{X}_n$ of n stationary dependent $\mathbb{R}^d$-valued random variables is (ϵ, α)-dense in $\mathbb{M} \subset \mathbb{R}^d$, for given $\epsilon > 0$ and $\alpha \in]0, 1[$, if

$$\mathbb{P}(d_H(\mathbb{X}_n, \mathbb{M}) \leq \epsilon) \geq 1 - \alpha.$$

If $\mathbb{X} := (X_i)_{i \in \mathbb{T}}$, with $\mathbb{T}$ being $\mathbb{Z}$, $\mathbb{N}$, or $\mathbb{N} \setminus \{0\}$, is a stationary sequence of $\mathbb{R}^d$-valued random variables, we say that $\mathbb{X}$ is *asymptotically dense* in $\mathbb{M} \subset \mathbb{R}^d$ if, for all positive ϵ sufficiently small and any $0 < \alpha < 1$, there exists a positive integer $n_0(\epsilon, \alpha)$ such that for every $n \geq n_0(\epsilon, \alpha)$, $\mathbb{X}_n = \{X_1, \ldots, X_n\}$ is (ϵ, α)-dense in $\mathbb{M}$.

The first undertaking of this paper is to identify sequences of dependent random vectors which are asymptotically dense in a compact support. In Sections 4 and 5 we treat explicitly a number of examples and show for all of these that the property of being asymptotically dense in the compact support holds by means of a key technical result, Proposition 3.1, which uses blocking techniques to give upper bounds for $\mathbb{P}(d_H(\mathbb{X}_n, \mathbb{M}) > \epsilon)$. Blocking techniques are very useful in the theory of limit theorems for stationary dependent random variables, and the underlying idea is to view and manipulate blocks as 'independent' clusters of dependent variables.

We summarize our first set of results into one main theorem. Given $\mathbb{X} := (X_i)_{i \in \mathbb{T}}$ a stationary sequence of $\mathbb{R}^d$-valued random variables, we denote by $\rho_m(\epsilon)$ the concentration quantity of the

https://doi.org/10.1017/apr.2024.4 Published online by Cambridge University Press

block $(X_1, \dots, X_m)$; that is, for $\epsilon > 0$,

$$\rho_m(\epsilon) := \inf_{x \in \mathbb{M}_{dm}} \mathbb{P}(\|(X_1, \cdots, X_m)^t - x\| \leq \epsilon),$$

where $\mathbb{M}_{dm}$ denotes the support of the vector transpose $(X_1, \cdots, X_m)^t$.

Theorem 1.1. *The following stationary sequences of $\mathbb{R}^d$-valued random variables are asymptotically dense in their common compact support:*

1. *Stationary m-dependent sequences such that for any $\epsilon > 0$, there exists a strictly positive constant κ_ϵ such that $\rho_{m+1}(\epsilon) \geq \kappa_\epsilon$. (See Proposition 4.1.)*
2. *Stationary m-approximable random variables on a compact set. These are stationary models that can be approximated by m-dependent stationary sequences (see Paragraph 4.0.2 and Proposition 4.2).*
3. *Stationary β-mixing sequences, with $(\beta_n)_n$ coefficients (see (4.6) for their definition), such that for some $\beta > 1$, and any $\epsilon > 0$,*

$$\lim_{m \to \infty} \rho_m(\epsilon) \frac{e^{m^\beta}}{m^{1+\beta}} = \infty, \quad \text{and} \quad \lim_{m \to \infty} \frac{e^{2m^\beta}}{m^2} \beta_m = 0.$$

   (See Proposition 4.3.)
4. *Stationary weakly dependent sequences, with $(\Psi(n))_n$ dependent coefficients (as introduced in (4.7)), such that for some $\beta > 1$ and any $\epsilon > 0$,*

$$\lim_{m \to \infty} \rho_m(\epsilon) \frac{e^{m^\beta}}{m^{1+\beta}} = \infty, \quad \text{and} \quad \lim_{m \to \infty} \frac{e^{2m^\beta}}{m^2} \Psi(m) = 0.$$

   (See Proposition 4.4.)
5. *Stationary Markov chains with an invariant measure μ and a suitable transition probability kernel (see Assumptions $(\mathcal{A}_1)$ and $(\mathcal{A}_2)$ of Section 5). (See Propositions 5.1 and 5.2.)*

Furthermore, and for each sequence $\mathbb{X}$ of random variables listed in Theorem 1.1, we give methods for finding a threshold $n_0(\epsilon, \alpha)$, sometimes with explicit formulae for it, such that $\mathbb{X}_n$ is (ϵ, α)-dense in the common support for $n \geq n_0(\epsilon, \alpha)$.

The next step is topological and consists in showing that when the Hausdorff distance between $\mathbb{X}_n$ and the support is sufficiently small, it is possible to reconstruct the support up to homotopy. We denote by $B(x, r)$ the closed ball in the Euclidean metric centered at x with radius $r > 0$, and we write $Y \xrightarrow{\simeq} X$ to mean that X deformation retracts onto Y with $Y \subset X$ (more precisely, this means that the identity map of X is homotopic to a retraction onto Y, leaving Y fixed during the homotopy).

Theorem 1.2. *Let $(X_i)_{i \in \mathbb{T}}$ be a stationary sequence of $\mathbb{R}^d$-valued random variables with compact support $\mathbb{M}$ having positive reach τ. Let $\epsilon \in \left(0, \frac{\tau}{2}\right)$, $r \in \left(\epsilon, \frac{\tau}{2}\right)$ and suppose that $\mathbb{X}_n$ is $(\frac{\epsilon}{2}, \alpha)$-dense in $\mathbb{M}$. Then*

$$\mathbb{P}\left(\mathbb{M} \xrightarrow{\simeq} \bigcup_{x \in \mathbb{X}_n} B(x, r) \right) \geq 1 - \alpha.$$

https://doi.org/10.1017/apr.2024.4 Published online by Cambridge University Press

The proof of this theorem is an immediate consequence of Definition 1.1 and a key reconstruction result proven in Section 2 (Theorem 2.2) which gives the same minimal conditions for recovering the homotopy type of the support $\mathbb{M}$ from a sample of points $\mathbb{X}_n$ in $\mathbb{M}$. Theorem 2.2 is 'deterministic' and should have wider application. The key geometric ideas behind this result are in [33], and in its extension in [39], as applied to the approximation of Riemannian submanifolds. To get Theorem 1.2, we weaken the regularity condition on the submanifolds from smooth to $C^{1,1}$, and in the hypersurface case we strengthen the bounds on the reach. This is then applied to thickenings of a positive-reach set M (see Section 2). It is important to contrast this result with earlier results in [29] (especially [29, Theorem 19]). There, the radii of the balls can be different. However, our Theorem 1.2 is simpler to state and easier to apply.

Having stated our main results, which are mainly of probabilistic and topological interest, we can say a few words about the statistical implications. In practice, the point-cloud data are realizations of random variables living in unknown support $\mathbb{M} \subset \mathbb{R}^d$. We then ask whether this support is a circle, a sphere, a torus, or a more complicated object. If we take sufficiently many points $\mathbb{X}_n$, our results tell us that the homology of $\mathbb{M}$ is the same as the homology of the union of balls around the data $\bigcup_{x\in\mathbb{X}_n} B(x, r)$, and this can be computed in general. The uniform radius r depends on $\mathbb{M}$ only through its reach, which is then the only quantity we need to estimate or to know a priori. Knowing the homology rules out many geometries for $\mathbb{M}$. Note that one may want to find ways to distinguish between a support that is a circle and one that is an annulus. However, conclusions of this sort are beyond the techniques of this paper.

1.1. Contents

We now give some more details about the content of the paper and how it is organized. We start by establishing, in Section 2, our result on homotopy reconstruction of a support from a point cloud in the deterministic case. Everything afterward is of probabilistic nature, with point clouds drawn from stationary random variables. In Sections 3 and 4 we state sufficient conditions for obtaining the asymptotically dense property, that is, conditions on concentrations and dependence coefficients under which $d_H(\mathbb{X}_n, \mathbb{M}) \leq \epsilon$ with large probability and for n large enough. More precisely, in Section 3 we give general upper bounds for $d_H(\mathbb{X}_n, \mathbb{M})$ using blocking techniques, i.e. by grouping the point cloud $\mathbb{X}_n$ into k_n blocks, each block with r_n points being considered as a single point in the appropriate Euclidean space of higher dimension. This is stated in Proposition 3.1, which is the key result of this paper, where the control of $d_H(\mathbb{X}_n, \mathbb{M})$ is reduced to the behavior of lower bounds of the concentration quantity of one block,

$$\rho_{r_n}(\epsilon) = \inf_{x\in\mathbb{M}_{dr_n}} \mathbb{P}(\|(X_1, \cdots, X_{r_n})^t - x\| \leq \epsilon), \tag{1.1}$$

and of

$$\inf_{x\in\mathbb{M}_{dr_n}} \mathbb{P}(\min_{1\leq i\leq k_n} \|(X_{(i-1)r_n+1}, \cdots, X_{ir_n})^t - x\| \leq \epsilon), \tag{1.2}$$

where, as before, $\mathbb{M}_{dr_n}$ is the support of the block $(X_1, \cdots, X_{r_n})^t$. Clearly, for independent random variables, a lower bound for (1.1) is directly connected to a lower bound for (1.2), but this is not the case for dependent random variables, and we need to control (1.1) and (1.2) separately. Section 4 gives our main examples of stationary sequences of $\mathbb{R}^d$-valued random variables having good convergence properties, under the Hausdorff metric, to the support. For each example we check that conditions needed for the control of (1.1) and (1.2) can be reduced

https://doi.org/10.1017/apr.2024.4 Published online by Cambridge University Press

to conditions on the concentration quantity $\rho_m(\epsilon)$ associated to the vector $(X_1, \cdots, X_m)^t$, for some fixed number of components $m \in \mathbb{N} \setminus \{0\}$. In particular, for mixing sequences, the control of $d_H(\mathbb{X}_n, \mathbb{M})$ is based on assumptions on the behavior of some lower bounds for this concentration quantity $\rho_m(\epsilon)$ in connection with the decay of the mixing dependence coefficients, as illustrated in Theorem 1.1. These lower bounds can be obtained by means of a condition similar to the so-called (a, b)-standard assumption (see for instance [7, 10, 11]) used in the case of i.i.d. sequences (i.e. when $k_n = n$ and $r_n = 1$). However, our results in Section 4 generalize the i.i.d. case without assuming the (a, b)-standard assumption (Subsection 4.1).

Section 5 gives explicit illustrations of our main results and techniques in the case of stationary Markov chains. For this model, the quantities in (1.1) and (1.2) can be controlled from the behavior of a positive measure ν defining the transition probability kernel of this Markov chain, in particular from the lower bounds of the concentration quantity $\nu(B(x, \epsilon) \cap \mathbb{M})$, for small ϵ and for $x \in \mathbb{M}$. The threshold $n_0(\epsilon, \alpha)$ can also be determined explicitly. As a key illustration, in Subsection 5.2 we study the Möbius Markov chain on the circle, where $\mathbb{M}$ is the unit circle and ν is the arc length measure on the unit circle. The conditions leading to a suitable control of (1.1) and (1.2) are checked with no further assumptions and the threshold $n_0(\epsilon, \alpha)$ is computed.

Section 6 gives an explicit simulation of a Möbius Markov chain studied in [26]. The intent here is to illustrate both the topological and probabilistic parts in an explicit situation. The simulation outcomes (Figures 4 and 5) are in agreement with the theoretical results thus obtained. Finally, all deferred proofs appear in Section 7.

2. A reconstruction result

Given a point cloud $S_n = \{x_1, \ldots, x_n\}$ on a metric space M, a standard problem is to reconstruct this space from the given distribution of points as n gets large (see the introduction). Various reconstruction processes in the literature are based on the nerve theorem. This basic but foundational result can be found in introductory books on algebraic topology ([23], chapter 4) and in most papers on manifold learning. This section takes a different route.

Below, let us write $B(x, r)$ (resp. $\mathring{B}(x, r)$) for the closed (resp. open) ball of radius r, centered at x. Starting with a point cloud $S_n = \{x_1, \ldots, x_n\} \subset M$, with M a compact subset of $\mathbb{R}^d$ with its Euclidean metric $\|\cdot\|$, we seek conditions on the radius r and on the distribution of the points of S_n ensuring that the union of balls $\bigcup_{i=1}^n B(x_i, r)$ deformation retracts onto M. The *r-offset* (or *r-thickening*, *r-dilation*, or *r-parallel set*, depending on the literature) of a closed set M is defined to be

$$M^{\oplus r} := \{p \in \mathbb{R}^d \mid d(p, M) := \inf_{x \in M} ||x - p|| \leq r\} = \bigcup_{x \in M} B(x, r).$$

Many of the existing theorems in homotopic and homological inference are about offsets. In terms of those, the Hausdorff distance d_H between two *closed* sets A and B is defined to be

$$d_H(A, B) = \inf_{r>0}\{A \subset B^{\oplus r}, B \subset A^{\oplus r}\} = \max\left(\sup_{x \in A} \inf_{y \in B} \|x - y\|, \sup_{x \in B} \inf_{y \in A} \|x - y\|\right) \tag{2.1}$$

(replacing inf and sup with min and max for compact sets). This is a 'coarse' metric in the sense that two closed spaces A and B can be very different topologically and yet be arbitrarily close in Hausdorff distance.

https://doi.org/10.1017/apr.2024.4 Published online by Cambridge University Press

We say that a subset $S \subset M$ is ϵ-dense (resp. strictly ϵ-dense) in M, for some $\epsilon > 0$, if $B(p, \epsilon) \cap S \neq \emptyset$ (resp. $\mathring{B}(p, \epsilon) \cap S \neq \emptyset$) for each $p \in M$. We have the following characterization.

Lemma 2.1. *Let $S \subset M$ be a closed subset. Then*

$$S \text{ is } \epsilon\text{-dense in } M \Longleftrightarrow M \subset S^{\oplus\epsilon} \iff d_H(S, M) \leq \epsilon.$$

Proof. When $S \subset M$, $d_H(S, M) = \inf\{r > 0 \mid M \subset S^{\oplus r}\}$. If S is ϵ-dense, any p in M is within ϵ of an $x \in S$, and so $M \subset S^{\oplus\epsilon}$, which implies that $d_H(S, M) \leq \epsilon$. The converse is immediate.

From now on, S will always mean a point cloud in M, that is, a finite collection of points. The following is a foundational result in the theory and is our starting point. □

Theorem 2.1. ([Proposition 3.1].) *Let M be a compact Riemannian submanifold of $\mathbb{R}^d$ with positive reach τ, and $S \subset M$ a strictly $\frac{\epsilon}{2}$-dense finite subset for $\epsilon < \sqrt{\frac{3}{5}}\tau$. Then for any $r \in [\epsilon, \sqrt{\frac{3}{5}}\tau[$, $M \xrightarrow{\simeq} \bigcup_{x\in S} \mathring{B}(x, r)$.*

Remark 2.1 Theorem 2.1 is a topological 'reconstruction' result which recovers the homotopy type of M from a finite sample. There are many reconstruction methods in the literature, which are too diverse to review here (see [3, 13, 29] and references therein). Reconstructions can be topological, meaning they recover the homotopy type or homology of the underlying manifold M, or they can be geometrical. We address only the topological aspect in this paper. In that regard, [29, Corollary 10] is attractive for its simplicity, as it proves a general reconstruction result for compact sets with positive reach by applying the nerve theorem to a cover by 'subspace balls' $\mathcal{U}_M = \{B(x_i, r) \cap M\}$. For Riemannian manifolds M, there is an alternative intrinsic geometric method for homotopy reconstruction based on 'geodesic balls'. Let $\rho_c > 0$ be a *convexity radius* for M. Such a radius has the property that around each $p \in M$, there is a 'geodesic ball' $B_g(p, \rho_c)$ which is convex, meaning that any two points in this neighborhood are joined by a unique geodesic in that neighborhood. These geodesic balls, and their non-empty intersections, are contractible. If $S_n = \{x_1, \ldots, x_n\}$ is a point cloud such that $\{B_g(x_i, \rho_c)\}$ is a cover of M, then the associated Čech complex is homotopy equivalent to M by the nerve theorem.

2.1. Positive reach

The notion of positive reach is foundational in convex geometry. As indicated in the introduction, the reach of a subset M is defined to be

$$\tau(M) := \sup\{r \geq 0 \mid \forall y \in M^{\oplus r} \ \exists!\ x \in M \text{ nearest to } y\}. \tag{2.2}$$

A PR set is any set M with $\tau(M) > 0$. Compact submanifolds are PR. Figure 1 gives an example of a PR set that is not a submanifold. The quintessential property of PR sets is the existence, for $0 < r < \tau$, of the 'unique closest point' projection

$$\pi_M : M^{\oplus r} \longrightarrow M, \ \ ||y - \pi_M(y)|| = d_H(y, M), \tag{2.3}$$

with $\pi_M(y)$ the unique nearest point to y in M. PR sets are necessarily closed, and thus compact if bounded.

As already indicated, we use the notation $Y \xrightarrow{\simeq} X$, if $Y \subset X$, to denote the fact that X deformation retracts onto Y. 'Thin-enough' offsets of PR sets deformation retract onto M.

Lemma 2.2. *Let M be a PR set with $\tau = \tau(M) > 0$. Then $M \xrightarrow{\simeq} M^{\oplus r}$ whenever $r < \tau$.*

https://doi.org/10.1017/apr.2024.4 Published online by Cambridge University Press

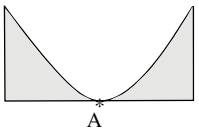

FIGURE 1. This space has positive reach τ in $\mathbb{R}^2$, but a neighborhood of point A indicates it is not a submanifold (with boundary).

Proof. This is immediate once we see that if $p \in M$ and $x \in \pi_M^{-1}(p) \subset M^{\oplus r}$, then the entire segment $[x, p]$ of $\mathbb{R}^d$ must be in $\pi_M^{-1}(p)$, so we can use the homotopy $F : M^{\oplus r} \times [0, 1] \to M^{\oplus r}$, $F(x, t) = (1 - t)x + t\pi_M(x)$, $t \in [0, 1]$, to get a (linear) deformation retraction, with M being fixed during the homotopy. □

The original interest in sets of positive reach lies in the fact that they have suitable small parallel neighborhoods with no self-intersections which allow one to compute their volume. This leads to a Steiner-type formula and a definition of curvature measures for these sets (see [38]). If M is a compact Riemannian submanifold in $\mathbb{R}^d$, as considered in [33], then $\tau(M)$ is positive and is the largest number having the property that the open normal bundle about M of radius r is smoothly embedded in $\mathbb{R}^d$ for every $r < \tau$. It is enough, however, for M to be C^2 to ensure that $\tau(M) > 0$ (see [38, Proposition 14]), and it is even enough for it to be $C^{1,1}$ in the case that M is a closed hypersurface (see [36, Theorem 1.3], which is an if-and-only-if statement). We recall the definition of $C^{1,1}$ (see [24, Definition 2.4.2]).

Definition 2.1. A closed manifold $M \subset \mathbb{R}^d$ is said to have $C^{1,1}$ boundary ∂M if for every $x_0 \in \partial M$ one can find a local open chart U of $x_0 \in \partial M$, and coordinates with the origin at x_0, such that $U \cap M = \{x \in M \mid x_1 \geq f(x')\}$, where $x' = (x_2, \ldots, x_d)$, f is C^1, and grad(f) is Lipschitz continuous.

Crucial to us in this section are the next two results. For general discussion, we refer to [17, 19] for Proposition 2.1, and [4, 36] for Proposition 2.2. Throughout, a manifold is assumed to be compact, and without boundary unless we specify the contrary. For $x \in \mathbb{R}^d$, let $d_M(x) = d(x, M) = \inf\{d(x, y), y \in M\}$ be the distance function to M. This function is 1-Lipschitz, and it is continuously differentiable when restricted to the interior of $M^{\oplus r} \setminus M$ if $r < \tau$ (see [17, Theorem 4.8]). Elementary point-set topology shows that the interior of the r-offset of M is $\text{int}(M^{\oplus r}) = \bigcup_{x \in M} \mathring{B}(x, r) = d_M^{-1}[0, r)$, and the topological boundary is $d_M^{-1}(r)$.

Proposition 2.1. ([17].) *Let $M \subset \mathbb{R}^d$ be compact of positive reach τ. For $0 < r < \tau$, $M^{\oplus r}$ is a compact manifold with $C^{1,1}$ boundary.*

We next describe tubular neighborhoods of a $C^{1,1}$-submanifold.

Proposition 2.2. *A closed submanifold N in $\mathbb{R}^d$, $d \geq 2$, has a tubular neighborhood 'foliated by orthogonal disks' if and only if it is $C^{1,1}$.*

Proof. This is a consequence of [36, Theorem 1.3], which proves that N is $C^{1,1}$ if and only if it has positive reach τ. A tubular neighborhood T (i.e. an embedding of the normal bundle extending the embedding of M) consists of all points at a distance strictly less than τ from N. This neighborhood has a unique nearest point projection $\pi_M : T \to M$. The orthogonal disks are the preimages of points in M under π_M. □

https://doi.org/10.1017/apr.2024.4 Published online by Cambridge University Press

Remark 2.2. A very informative discussion about the above is on MathOverflow [31], and the point is this. In the C^1 case, the choice of the (unit, outer) normal vector at every point of N is a continuous function (this is by definition the Gauss map). In fact if N is C^k, then the choice of a normal $N \to \mathbb{R}^n$ is C^{k-1} (see [4, Lemma 4.6.18]). If we have C^1-regularity but not $C^{1,1}$, it could happen that the normals intersect arbitrarily close to the hypersurface, in which case the reach is indeed 0. A good example to keep in mind, which we owe to S. Scholtes (private communication), is the graph of the real-valued function which is 0 for $x \leq 0$ and $x^{3/2}$ for $x \geq 0$. This function is $C^{1,1/2}$, not $C^{1,1}$, and one observes that near 0, the normals intersect arbitrarily close to the curve.

If M is a compact PR set, its offset $M^{\oplus r}$ is also compact and PR for $r < \tau$, with reach $\tau - r$, where τ is the reach of M. This assertion is not entirely obvious, since, in general, the reach is not always well-behaved for nested compact sets. By this we mean that if (K_2, K_1) is a pair of nested compact sets in $\mathbb{R}^d$, $K_1 \subset K_2$, then both cases $\tau_1 < \tau_2$ or $\tau_2 < \tau_1$ can occur, where τ_i is the reach of K_i. As an example of the former case, take K_1 to be the circle and K_2 to be the closed disk; for the latter case, take K_1 to be a point in a finite-reach K_2. The case of $(K_2, K_1) = (M^{\oplus r}, M)$ is therefore special.

Lemma 2.3. *Let $M \subset \mathbb{R}^d$ be a compact PR set with reach τ, and $0 \leq r < \tau$; then $M^{\oplus r}$ has positive reach with $\tau(M^{\oplus r}) = \tau - r > 0$.*

Proof. Essentially, the point is that any ray from $y \notin M^{\oplus r}$ to M must cut the boundary $\partial M^{\oplus r}$ at a point lying at a distance of r to M. Suppose $r > 0$, so that $M^{\oplus r}$ is a codimension-0 manifold with boundary $\partial M^{\oplus r}$ in $\mathbb{R}^d$. Write τ_r for its reach. We will first prove that if y in the complement of $M^{\oplus r}$ has a unique projection onto M, then necessarily it has a unique projection onto $M^{\oplus r}$ (this will prove that $\tau_r > 0$ and that $\tau - r \leq \tau_r$). Reciprocally, we will argue that if y has a unique projection onto M, then it also has a unique projection onto $M^{\oplus r}$.

To prove the first claim, write $y_1 = [y, \pi_M(y)] \cap \partial M^{\oplus r}$. We claim that y_1 is the unique closest point to y in $M^{\oplus r}$. Indeed, if there is z_1 on that boundary that is closer to y, then

$$d(y, \pi_M(z_1)) \leq d(y, z_1) + d(z_1, \pi_M(z_1)) = d(y, z_1) + r \leq d(y, y_1) + d(y_1, M) = d(y, \pi_M(y)),$$

and so $d(y, \pi_M(z_1)) = d(y, \pi_M(y))$ (since $d(y, \pi_M(y))$ is smallest distance from y to M), and by uniqueness, $\pi_M(z_1) = \pi_M(y)$. This implies that $d(y, z_1) + d(z_1, M) = d(y, M)$, and so $y, z_1, \pi_M(y)$ are aligned. This can only happen if $y_1 = z_1$.

Suppose now that y has a unique projection onto $M^{\oplus r}$, which we label y'. We can check that it also has a unique projection onto M. Let z be that projection. By a similar argument as previously, z must be $\pi_M(y')$ (so unique) and y, y',z are aligned. This shows reciprocally that $\tau_r + r \leq \tau$.

The above arguments show that $\tau = \tau_r + r$, and in fact they can be used to show in this case that $M^{\oplus r'} = (M^{\oplus r})^{\oplus (r'-r)}$ for all $r \leq r' < \tau$. □

2.2. Manifolds with boundary

In order to apply our ideas to PR sets, we need to extend Theorem 2.1 from closed Riemannian submanifolds to submanifolds with boundary. Note that the reach of ∂M (manifold boundary) and M are not comparable in general. Indeed, take M to be the $y = \sin(x)$ curve on $[0, \pi]$, with boundary the endpoints. Then $\tau(M) < \tau(\partial M)$. Take now a closed disk M in $\mathbb{R}^2$. Then $\tau(\partial M) < \tau(M) = \infty$. If M is of codimension 0, then $\tau(\partial M) \leq \tau(M)$ always.

https://doi.org/10.1017/apr.2024.4 Published online by Cambridge University Press

In [39], the authors managed to extend Theorem 2.1 to smooth submanifolds with boundary and showed that in this case the bound $\sqrt{\frac{3}{5}}\tau$ in Theorem 2.1 can be replaced by $\frac{\delta}{2}$, where $\delta = \min(\tau(M), \tau(\partial M))$. We revisit this result in the codimension-0 case and establish the following 'twice the density, half the reach' criterion.

Proposition 2.3. *Let M be a compact codimension-0 submanifold of $\mathbb{R}^d$ with $C^{1,1}$ boundary and having positive reach $\tau = \tau(M) > 0$, and let $S \subset M$ be an $\frac{\epsilon}{2}$-dense finite subset with $\epsilon < \frac{\tau}{2}$. Then for any r such that $\epsilon \leq r < \frac{\tau}{2}$, $M \xrightarrow{\simeq} \bigcup_{x\in S} B(x, r)$.*

Notice that M need not be connected. Notice also that we have weakened the regularity on ∂M from smooth to $C^{1,1}$. According to [36, Theorem 1.3] (see also [17, Remark 4.20]), this condition is enough to ensure that $\tau(\partial M) > 0$. Finally, notice that we use closed balls in our statement, and that they may have radius larger than ϵ, but not exceeding $\frac{\tau}{2}$.

Proof. The proof is an adaptation of Lemma 4.1 of [33] and Lemma 4.3 of [39] for smooth submanifolds. For completeness, we will reconstruct the part of the argument that we need.

Firstly, since the reach of a disjoint finite union of PR sets is the least of their reaches and their pairwise distances, we can assume without loss of generality that M is connected from the start. Let M be connected of codimension 0. Then its boundary is a connected codimension-1 closed submanifold (i.e. a closed hypersurface). It divides Euclidean space into two regions: M and its complement. We let τ^- denote the reach of a component of ∂M in M (the interior region), and τ^+ its reach within the open (exterior) region. Clearly $\tau := \tau(M) = \tau^+$.

Now ∂M is $C^{1,1}$ by hypothesis, and being a closed hypersurface, it is necessarily orientable. It has a continuous normal vector field into the exterior region, defining a (trivial) $\mathbb{R}_+$-bundle $T(\partial M)^+$ of $T(\partial M)$. We write $T_p^{\perp,+}(\partial M)$ for the fiber at $p \in \partial M$, which is a half-line extending into the exterior region, perpendicular to $T_p(\partial M)$. Linear deformation retraction along this direction as in Lemma 2.2, keeping M fixed, shows that $M \xrightarrow{\simeq} M^{\oplus r}$ as long as $r < \tau^+ = \tau$ (where normal directions never intersect). We have that

$$M^{\oplus r} = \bigcup_{x\in M} B(x, r) \simeq M, \quad r < \tau.$$

We want to show that this retraction of $M^{\oplus r}$ onto M (along fibers of $T^+(\partial M)$) restricts to a deformation retraction onto M of the middle space $S^{\oplus r}$ in the sequence of inclusions below:

$$M \subset S^{\oplus r} = \bigcup_{x\in S} B(x, r) \subset M^{\oplus r}, \quad \epsilon \leq r < \frac{\tau}{2}.$$

That is, we only take the union of balls centered at points of S. This covers M since $r \geq \frac{\epsilon}{2}$ and any point of M is within distance $\epsilon/2$ of S. Let us see how the deformation retraction of the bigger space $M^{\oplus r}$ onto M may fail to restrict to a retraction on $S^{\oplus r}$: let $v \in T_p^{\perp,+}(\partial M)$, and suppose $v \in B(q, r)$, with $q \in S$ but $q \notin B(p, r)$. So the line segment $[v, p]$ is not in the ball $B(q, r)$, and the linear retraction will leave that ball eventually. This, however, will not cause a problem as long as the segment falls in another ball and does not leave the entire union $\bigcup_{x\in S} B(x, r)$. This happens if both v and p are in some other ball $B(x, r)$, $x \in S$ (because balls are convex). By the density condition, we can also demand that x be at a distance of at most $\frac{\epsilon}{2}$ from p. To recapitulate, for every such p, v, by picking an $x \in S$ within a distance of $\frac{\epsilon}{2}$ from p and a distance of r from v, we guarantee that the deformation retraction of $M^{\oplus r}$ restricts to $S^{\oplus r}$ (see Figure 2).

https://doi.org/10.1017/apr.2024.4 Published online by Cambridge University Press

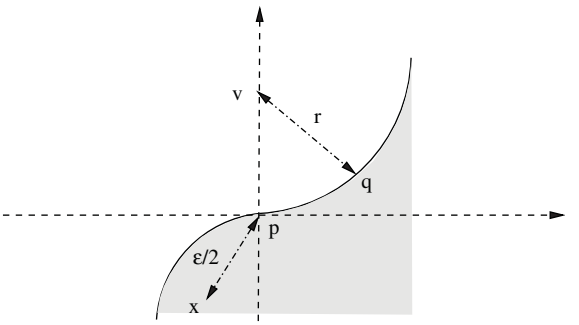

FIGURE 2. $M \subset \mathbb{R}^d$ is represented by the shaded area, $p, q \in \partial M$, and $x, q \in S$. The points q, p are on a circle tangent to $T_p(M)$, having radius τ and center on the vertical dashed line representing the normal direction $T_p^{\perp,+}(M)$, pointing into the exterior region, while x is anywhere in $M \cap S$ at a distance of at most $\frac{\epsilon}{2}$ from p. An extreme disposition of such points (meaning when v is as far as possible from x) happens when v, p, x are aligned. This figure is the analog of Figure 2 of [33] and Figure 1 of [39].

With the above target in mind, consider the following configuration of points: $p \in \partial M$, $v \in T_p^{\perp,+}(\partial M) \cap B(p, \tau)$, $q \in S$, and $v \in B(q, r)$ but $p \notin B(q, r)$. How far can v be from p, among all choices of such points q? The answer can be extracted from the key Lemma 4.1 of [33]. The worst-case scenario, corresponding to when v would be farthest from p, is when q and p lie on the circle of radius τ, with center in $T_p^{\perp,+}$ as in Figure 2, and p,q,v make up an isosceles triangle with $|p - q| = |q - v| = r$. Lemma 4.1 of [33] applied to this situation gives that $d(v, p) < \dfrac{r^2}{\tau^+} = \dfrac{r^2}{\tau} < \tau$.

Next, we look for an $x \in B(p, \frac{\epsilon}{2}) \cap S \neq \emptyset$ which is within distance r of v. By the triangle inequality, $d(x, v) \leq d(x, p) + d(p, v) \leq \dfrac{\epsilon}{2} + \dfrac{r^2}{\tau}$, and so in order for $v \in B(x, r)$, it is enough to require that

$$\frac{\epsilon}{2} + r^2\tau < r \iff r^2 - r\tau + \frac{\epsilon\tau}{2} < 0. \tag{2.4}$$

Any value of r between the roots of the polynomial on the right-hand side does the job, and it is immediate that $r \in \left[\epsilon, \frac{\tau}{2}\right[$ satisfies this condition. The proof is complete.

Note that in the case of Theorem 2.1 in [33], that is, when one is considering M that is closed (i.e. with no boundary), the point x to be chosen cannot simply be 'anywhere' around $p \in M$, as in Figure 2, but must lie on M (which would be the boundary in that figure), and thus the authors get a different bound on r. □

We finally come to the proof of the main reconstruction result of this section, which yields Theorem 1.2 as a consequence. We thank the referee for suggesting this statement, which is simpler than the one we originally gave.

Theorem 2.2. *Let M be a compact space in $\mathbb{R}^d$ with positive reach τ, let $\epsilon \in \left(0, \frac{\tau}{2}\right)$, and let $r \in \left(\epsilon, \frac{\tau}{2}\right)$. If $S \subset M$ is $\frac{\epsilon}{2}$-dense, then $M \xrightarrow{\simeq} \bigcup_{x \in S} B(x, r)$.*

Proof. Notice that by Proposition 2.3, the result is true if M is a codimension-0 submanifold whose boundary is $C^{1,1}$ of reach $\tau > 0$ (let us refer to this submanifold as a 'good object'). The

https://doi.org/10.1017/apr.2024.4 Published online by Cambridge University Press

idea now is very simple and relies on the fact that, if M itself is not such a good object but is of positive reach, then any slight thickening of it will be a good object.

Assume that S is $\epsilon/2$-dense in M, with $\epsilon < \frac{\tau}{2}$. We first collect the following facts:

1. For $0 < \delta < \tau$, $M^{\oplus\delta}$ deformation retracts onto M (Lemma 2.2).
2. For $0 < \delta < \tau$, S is $\left(\frac{\epsilon}{2} + \delta\right)$-dense in $M^{\oplus\delta} \supset M$.
3. The offset $M^{\oplus\delta}$ is a codimension-0 submanifold of $\mathbb{R}^d$, with $C^{1,1}$ boundary (Proposition 2.1). Its reach is $\tau' = \tau - \delta$ (Lemma 2.3).

Assume that $\delta < \dfrac{\tau - 2\epsilon}{5}$. This is equivalent to $\epsilon + 2\delta < \frac{\tau-\delta}{2}$; that is, twice the density of S in $M^{\oplus\delta}$ is less than half the reach of $M^{\oplus\delta}$. By Proposition 2.3, for all r that we can insert between these two numbers, we get a homotopy reconstruction of $M^{\oplus\delta}$; more precisely,

$$\epsilon + 2\delta \le r < \frac{\tau - \delta}{2} \implies \bigcup_{x\in S} B(x, r) \simeq M^{\oplus\delta}. \tag{2.5}$$

Returning to the hypotheses of Theorem 2.2, choose any r such that $\epsilon < r < \frac{\tau}{2}$. Pick δ such that $0 < \delta < \min\left\{\dfrac{r-\epsilon}{2}, \tau - 2r\right\}$. For this δ, $\epsilon + 2\delta \le r$ and $r < \frac{\tau-\delta}{2}$, and thus by (2.5), $\bigcup_{x\in S} B(x, r)$ deformation retracts onto $M^{\oplus\delta}$. Since the latter deformation retracts onto M, the composition of both retractions shows that $M \xrightarrow{\simeq} \bigcup_{x\in S} B(x, r)$. The proof is complete. □

3. Blocking techniques and upper bounds for the Hausdorff distance

In this section we state and prove the main technical result of this paper. This is given by Proposition 3.1 below, which is general and of independent interest. It is based on blocking techniques as well as a useful geometrical result, proven in [33], relating the minimal covering number of a compact set by closed balls to the maximal length of a chain of points whose pairwise distances are bounded below.

Let $(X_i)_{i\in\mathbb{T}}$ (where $\mathbb{T}$ is either $\mathbb{Z}$, $\mathbb{N}$, or $\mathbb{N}\setminus\{0\}$) be a stationary sequence of $\mathbb{R}^d$-valued random variables. Let P be the distribution of X_1. Suppose that P is supported on a compact set $\mathbb{M}$ of $\mathbb{R}^d$, i.e. $\mathbb{M} := \operatorname{supp}(X_j)$ is the smallest closed set carrying the mass of P:

$$\mathbb{M} = \bigcap_{C\subset\mathbb{R}^d,\ P(\overline{C})=1} \overline{C}, \tag{3.1}$$

where $\overline{C}$ means the closure of the set C in Euclidean space. Recall that $\mathbb{X}_n = \{X_1, \cdots, X_n\}$ and this is viewed as a subset of $\mathbb{R}^d$. Throughout, we will be working with the Hausdorff distance d_H (2.1). Note that $d_H(\{x\}, \{y\}) = ||x - y||$ (in the Euclidean distance) if x, y are points. Note that the distance of a point y to a closed set A is $d(y, A) = \inf_{x\in A} ||x - y||$, while its Hausdorff distance to A is $d_H(y, A) = \sup_{x\in A} ||x - y||$. This explains in part why this metric is very sensitive to outliers (see [32]) and to noisy phenomena.

We wish to give upper bounds for $\mathbb{P}(d_H(\mathbb{X}_n, \mathbb{M}) > \epsilon)$ via a blocking technique. Let k and r be two positive integers such that $kr \le n$. For $1 \le i \le k$, define the random vector $Y_{i,r}$ in $\mathbb{R}^{dr}$ by $Y_{i,r} = (X_{(i-1)r+1}, \cdots, X_{ir})^t$. Let

$$\mathbb{Y}_k = \{Y_{1,r}, \cdots, Y_{k,r}\}$$

https://doi.org/10.1017/apr.2024.4 Published online by Cambridge University Press

be a subset in $\mathbb{R}^{dr}$ of k stationary random vectors which are not necessarily independent. The support $\mathbb{M}_{dr}$ of the vector $Y_{1,r}$ is included in $\mathbb{M} \times \cdots \times \mathbb{M}$ (r times), and since, by definition, $\mathbb{M}_{dr}$ is a closed set, it is necessarily compact in $\mathbb{R}^{dr}$. As we now show, it is possible to reduce the behavior of $d_H(\mathbb{X}_n, \mathbb{M})$ to that of the sequence of vectors $(Y_{i,r})_{1\le i\le k}$ for any k and r for which $kr \le n$ and under only the assumption of stationarity of $(X_i)_{i\in\mathbb{T}}$.

Proposition 3.1. *With $\epsilon > 0$, and with k and r any positive integers such that $kr \le n$, it holds that*

$$\mathbb{P}(d_H(\mathbb{X}_n, \mathbb{M}) > \epsilon) \le \mathbb{P}(d_H(\mathbb{Y}_k, \mathbb{M}_{dr}) > \epsilon) \le \frac{\sup_{x\in\mathbb{M}_{dr}} \mathbb{P}\big(\min_{1\le i\le k} \|Y_{i,r} - x\| > \epsilon/2\big)}{1 - \sup_{x\in\mathbb{M}_{dr}} \mathbb{P}\big(\|Y_{1,r} - x\| > \epsilon/4\big)}.$$

Proof. Since $\mathbb{P}(\mathbb{Y}_k \subset \mathbb{M}_{dr}) = 1$, we have almost surely (a.s.)

$$d_H(\mathbb{Y}_k, \mathbb{M}_{dr}) = \sup_{x\in\mathbb{M}_{dr}} \min_{1\le j\le k} \|Y_{j,r} - x\|. \tag{3.2}$$

Since $\mathbb{M}_{dr}$ is compact, there exists a finite set $\mathcal{C}_N = \{c_1, \cdots, c_N\} \subset \mathbb{M}_{dr} \subset \mathbb{R}^{dr}$ of centers of balls forming a minimal ϵ-covering set for $\mathbb{M}_{dr}$, so that, for a fixed $x \in \mathbb{M}_{dr}$, there exists $c_i \in \mathcal{C}_N \subset \mathbb{M}_{dr}$ such that

$$\|x - c_i\| \le \epsilon.$$

Hence,

$$\|Y_{j,r} - x\| \le \|Y_{j,r} - c_i\| + \|c_i - x\| \le \|Y_{j,r} - c_i\| + \epsilon.$$

Consequently, for any $x \in \mathbb{M}_{dr}$,

$$\min_{1\le j\le k} \|Y_{j,r} - x\| \le \min_{1\le j\le k} \|Y_{j,r} - c_i\| + \epsilon \le \max_{1\le i\le N} \min_{1\le j\le k} \|Y_{j,r} - c_i\| + \epsilon$$

and

$$\sup_{x\in\mathbb{M}_{dr}} \min_{1\le j\le k} \|Y_{j,r} - x\| \le \max_{1\le i\le N} \min_{1\le j\le k} \|Y_{j,r} - c_i\| + \epsilon.$$

Hence,

$$\mathbb{P}\left(\sup_{x\in\mathbb{M}_{dr}} \min_{1\le j\le k} \|Y_{j,r} - x\| \ge 2\epsilon\right) \le \mathbb{P}\left(\max_{1\le i\le N} \min_{1\le j\le k} \|Y_{j,r} - c_i\| \ge \epsilon\right)$$

$$\le N \max_{1\le i\le N} \mathbb{P}\left(\min_{1\le j\le k} \|Y_{j,r} - c_i\| \ge \epsilon\right) \le N \sup_{x\in\mathbb{M}_{dr}} \mathbb{P}\left(\min_{1\le j\le k} \|Y_{j,r} - x\| \ge \epsilon\right). \tag{3.3}$$

We now have to bound N. For this we use [33, Lemma 5.2] (as was done in [16]), to get

$$N \le \left(\inf_{x\in\mathbb{M}_{rd}} \mathbb{P}(\|Y_{1,r} - x\| \le \epsilon/2)\right)^{-1} = \left(1 - \sup_{x\in\mathbb{M}_{rd}} \mathbb{P}(\|Y_{1,r} - x\| > \epsilon/2)\right)^{-1}. \tag{3.4}$$

Hence, by (3.2) together with (3.3) and (3.4),

$$\mathbb{P}(d_H(\mathbb{Y}_k, \mathbb{M}_{dr}) > 2\epsilon) \tag{3.5}$$

$$\le \left(1 - \sup_{x\in\mathbb{M}_{rd}} \mathbb{P}(\|Y_{1,r} - x\| > \epsilon/2)\right)^{-1} \sup_{x\in\mathbb{M}_{rd}} \mathbb{P}\left(\min_{1\le j\le k} \|Y_{j,r} - x\| \ge \epsilon\right).$$

https://doi.org/10.1017/apr.2024.4 Published online by Cambridge University Press

Thanks to (3.5), the proof of this proposition is complete if we prove that

$$\mathbb{P}(d_H(\mathbb{X}_n, \mathbb{M}) > \epsilon) \leq \mathbb{P}(d_H(\mathbb{Y}_k, \mathbb{M}_{dr}) > \epsilon)\,. \tag{3.6}$$

Recall that $\mathbb{P}(\mathbb{X}_n \subset \mathbb{M}) = 1$, so that $d_H(\mathbb{X}_n, \mathbb{M}) = \sup_{x\in\mathbb{M}} \min_{1\leq j\leq n} \|X_j - x\|$, and, since $kr \leq n$,

$$d_H(\mathbb{X}_n, \mathbb{M}) = \sup_{x\in\mathbb{M}} \min_{1\leq j\leq n} \|X_j - x\| \leq \sup_{x\in\mathbb{M}} \min_{1\leq j\leq kr} \|X_j - x\| = d_H(\mathbb{X}_{kr}, \mathbb{M}).$$

From this we deduce that

$$\mathbb{P}(d_H(\mathbb{X}_n, \mathbb{M}) > \epsilon) \leq \mathbb{P}(d_H(\mathbb{X}_{kr}, \mathbb{M}) > \epsilon)\,. \tag{3.7}$$

It finally remains to prove that

$$\mathbb{P}(d_H(\mathbb{X}_{kr}, \mathbb{M}) > \epsilon) \leq \mathbb{P}(d_H(\mathbb{Y}_k, \mathbb{M}_{dr}) > \epsilon)\,. \tag{3.8}$$

For this, let $X_j \in \mathbb{X}_{kr}$ and $x \in \mathbb{M}$. Then there exist l and i such that X_j is the lth component of the vector $Y_{i,r}$. We claim also that there exists $\tilde{x} \in \mathbb{M}_{dr}$ such that x is the lth component of the vector $\tilde{x}$. In fact, let $\pi_l : \mathbb{R}^{dr} \to \mathbb{R}^d$ be the projection onto the lth factor. It follows from an elementary property of the support, by the continuity of π_l and the closure of $\mathbb{M}_{dr}$, that

$$\mathbb{M} = \operatorname{supp}(X_j) = \overline{\pi_l(\operatorname{supp}(Y_{i,r}))} = \overline{\pi_l(\mathbb{M}_{dr})} = \pi_l(\mathbb{M}_{dr}),$$

where $\overline{A}$ denotes, as before, the closure of the set A. So, in particular, any $x \in \mathbb{M}$ is $x = \pi_l(\tilde{x})$ for some $\tilde{x} \in \mathbb{M}_{dr}$. From this, we deduce that, for any $X_j \in \mathbb{X}_{kr}$ and $x \in \mathbb{M}$, there exist $1 \leq i \leq k$ and $\tilde{x} \in \mathbb{M}_{dr}$ such that, a.s.,

$$\|X_j - x\| \leq \|Y_{i,r} - \tilde{x}\|.$$

Hence,

$$\inf_{X_j\in\mathbb{X}_{kr}} \|X_j - x\| \leq \inf_{Y_{i,r}\in\mathbb{Y}_k} \|Y_{i,r} - \tilde{x}\| \leq d_H(\mathbb{Y}_k, \mathbb{M}_{dr}).$$

Consequently, since $\mathbb{P}(\mathbb{X}_{kr} \subset \mathbb{M}) = 1$,

$$d_H(\mathbb{X}_{kr}, \mathbb{M}) = \sup_{x\in\mathbb{M}} \inf_{X_j\in\mathbb{X}_{kr}} \|X_j - x\| \leq d_H(\mathbb{Y}_k, \mathbb{M}_{dr}).$$

From this we get (3.8). Now (3.8) together with (3.7) proves (3.6). The proof of the proposition is complete. □

4. Asymptotically dense sequences of random variables

As indicated in the introduction, our main goal is to find conditions under which a sequence $\mathbb{X}$ is asymptotically dense in the common support (see Definition 1.1). In this section, we give conditions and several examples of dependent random variables for which this is the case. This property is established every time by means of Proposition 3.1 applied with suitable choices of subsequences k and r of n, and for all these examples, it holds that for any $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}(d_H(\mathbb{Y}_k, \mathbb{M}_{dr}) > \epsilon) = \lim_{n\to\infty} \mathbb{P}(d_H(\mathbb{X}_n, \mathbb{M}) > \epsilon) = 0.$$

All proofs of the propositions listed in this section appear in Section 7.

https://doi.org/10.1017/apr.2024.4 Published online by Cambridge University Press

4.0.1. *Stationary m-dependent sequence on a compact set.* Recall that the sequence $(X_i)_{i\in\mathbb{T}}$ is m-dependent for some $m \geq 0$ if the two σ-fields $\sigma(X_i,\ \ i \leq k)$ and $\sigma(X_i, i \geq k+m+1)$ are independent for every k. In particular, 0-dependent is the same as independent.

Example 4.1. *(An m-dependent sequence.) Let $(T_i)_{i\in\mathbb{N}}$ be a sequence of i.i.d. random variables with values in $\mathbb{R}^d$. Let h be a real-valued function defined on $\mathbb{R}^{dm}$. The stationary sequence $(X_n)_{n\in\mathbb{N}}$ defined by $X_n = h(T_n, T_{n+1}, \cdots, T_{n+m})$ is a stationary sequence of m-dependent random variables.*

For $m \in \mathbb{N}\setminus\{0\}$, $\epsilon > 0$, and $Y_{1,m} = (X_1, \cdots, X_m)^t$, as in the introduction, define the concentration coefficient of the vector $Y_{1,m}$ by

$$\rho_m(\epsilon) = \inf_{x\in\mathbb{M}_{dm}} \mathbb{P}\big(\|Y_{1,m} - x\| \leq \epsilon\big)\,. \tag{4.1}$$

The following proposition gives conditions on $\rho_m(\epsilon)$ under which the asymptotically dense property evoked in Definition 1.1 is satisfied.

Proposition 4.1. *Let $(X_i)_{i\in\mathbb{T}}$ be a stationary sequence of m-dependent $\mathbb{R}^d$-valued random vectors. Suppose that X_1 has compact support $\mathbb{M}$. Let $\epsilon_0 > 0$ be fixed. Suppose that for any $0 < \epsilon < \epsilon_0$, there exists a strictly positive constant κ_ϵ such that*

$$\rho_{m+1}(\epsilon) \geq \kappa_\epsilon.$$

Then it holds for any $0 < \epsilon < \epsilon_0$ and any $n \geq m+1$ that

$$\mathbb{P}(d_H(\mathbb{X}_n, \mathbb{M}) > \epsilon) \leq \frac{(1-\kappa_{\frac{\epsilon}{2}})^{[\frac{1}{2}[\frac{n}{m+1}]]}}{\kappa_{\frac{\epsilon}{4}}},$$

where $[\,\cdot\,]$ denotes the integer part. Consequently, for any $\alpha \in]0,1[$ and any $n \geq n_0(\epsilon,\alpha)$, where

$$n_0(\epsilon,\alpha) = \frac{2(m+1)}{\kappa_{\frac{\epsilon}{2}}}\left(\log\left(\frac{1}{\alpha}\right) + \log\left(\frac{1}{\kappa_{\frac{\epsilon}{4}}}\right)\right) + 3(m+1),$$

we have $d_H(\mathbb{X}_n, \mathbb{M}) \leq \epsilon$ with probability at least $1-\alpha$.

The requirements of Proposition 4.1 prove that the sequence $(X_n)_{n\in\mathbb{T}}$ is asymptotically dense in $\mathbb{M}$ with threshold $n_0(\epsilon,\alpha)$ as above.

4.0.2. *Stationary m-approximable random variables on a compact set.* In this section we discuss, in the spirit of [25], some examples of stationary compactly supported random variables $(X_n)_{n\in\mathbb{Z}}$ that can be approximated by m-dependent stationary sequences. More precisely, the article [25] introduced the notion of an L^p-m-approximable sequence. This notion is related to m-dependence (see [25, Definition 2.1]) and is different from mixing (see Paragraph 4.0.3 below for a definition of mixing). The idea is to construct, for $m \in \mathbb{N}$, a stationary sequence $(X_n^{(m)})_{n\in\mathbb{Z}}$ that is m-dependent and compactly supported, for which the Hausdorff distance between the two sets $\mathbb{X}_n$ and $\mathbb{X}_n^{(m)} := \{X_1^{(m)}, \cdots, X_n^{(m)}\}$ is suitably controlled. In our case, it is not necessary that X_1 and $X_1^{(m)}$ have the same distribution; rather, what we need is that X_1 and X_1^m have the same compact support. This will give us more choices for the construction of the sequence $(X_n^{(m)})_{n\in\mathbb{Z}}$, which can be obtained by the method of coupling or by a truncation argument (see [25] for more details). For our purpose, we shall use a truncation.

https://doi.org/10.1017/apr.2024.4 Published online by Cambridge University Press

More precisely, we will consider the sequence

$$X_n = f(\epsilon_n, \epsilon_{n-1}, \cdots), \tag{4.2}$$

where $(\epsilon_i)_{i\in\mathbb{Z}}$ is an i.i.d. sequence with values in some measurable space S and f is a real bounded function defined on S^∞. The sequence $(X_n^{(m)})_{n\in\mathbb{Z}}$ constructed from $(X_n)_{n\in\mathbb{Z}}$ by truncation is

$$X_n^{(m)} = f(\epsilon_n, \cdots, \epsilon_{n-m}, 0, \cdots). \tag{4.3}$$

Clearly $(X_n^{(m)})_{n\in\mathbb{Z}}$ is a stationary, m-dependent sequence and has the same compact support as $(X_n)_{n\in\mathbb{Z}}$ as soon as f is bounded. We will thus assume that f is bounded; that is, $\|f\|_\infty = \sup_{x\in S^\infty} |f(x)| < \infty$. As before, $\mathbb{M}$ will be the common support of these two sequences.

Now we need an additional assumption on f in order to ensure good control of the Hausdorff distance between the two sets $\mathbb{X}_n$ and $\mathbb{X}_n^{(m)}$. We suppose that f is a real-valued bounded function and that it satisfies the following Lipschitz-type assumption (stated in [25]): there exists a decreasing sequence $(c_m)_{m\in\mathbb{N}}$, tending to 0 as m tends to infinity, such that

$$|f(a_{m+1}, \cdots, a_1, x_0, \cdots) - f(a_{m+1}, \cdots, a_1, y_0, \cdots)| \le c_m |f(x_0, x_{-1}, \cdots) - f(y_0, y_{-1}, \cdots)|, \tag{4.4}$$

for any numbers $a_l, x_i, y_i \in S$, $l \in \{1, m+1\}$, and $i \le 0$. This assumption is satisfied, for instance, by some autoregressive models.

The next lemma proves that the truncated sequence $(X_n^{(m)})_{n\in\mathbb{Z}}$ is a Hausdorff approximation of the original sequence $(X_n)_{n\in\mathbb{Z}}$.

Lemma 4.1. *Let $\epsilon > 0$ be fixed, let $(\epsilon_i)_{i\in\mathbb{Z}}$ be an i.i.d. sequence, let f be a bounded function satisfying (4.4), and let $(X_n)_{n\in\mathbb{Z}}$ and $(X_n^{(m)})_{n\in\mathbb{Z}}$ be the associated sequences as in (4.2) and (4.3), respectively. Let $m \in \mathbb{N}$ be such that*

$$2c_m \|f\|_\infty < \epsilon,$$

where $\|f\|_\infty$ is the supremum of f. Then $d_H(\mathbb{X}_n^{(m)}, \mathbb{X}_n) < \epsilon$, a.s.

In view of Lemma 4.1, the condition $\lim_{m\to\infty} c_m = 0$ is enough to approximate, in the Hausdorff sense, the sequence $(X_n)_{n\in\mathbb{Z}}$ by the truncated sequence $(X_n^{(m)})_{n\in\mathbb{Z}}$.

Proof. We recall that

$$d_H(\mathbb{X}_n^{(m)}, \mathbb{X}_n) = \max\left(\max_{1\le i\le n} \min_{1\le j\le n} |X_i - X_j^{(m)}|, \max_{1\le i\le n} \min_{1\le j\le n} |X_j - X_i^{(m)}| \right).$$

Hence,

$$\begin{aligned}
&\mathbb{P}(d_H(\mathbb{X}_n^{(m)}, \mathbb{X}_n) \ge \epsilon) \\
&\le \mathbb{P}\Big(\max_{1\le i\le n} \min_{1\le j\le n} |X_i - X_j^{(m)}| \ge \epsilon \Big) + \mathbb{P}\Big(\max_{1\le i\le n} \min_{1\le j\le n} |X_j - X_i^{(m)}| \ge \epsilon \Big) \\
&\le 2n \max_{1\le i\le n} \mathbb{P}(|X_i - X_i^{(m)}| \ge \epsilon).
\end{aligned}$$

Now,

$$|X_i - X_i^{(m)}| \le c_m |f(\epsilon_{n-m-1}, \epsilon_{n-m-2}, \cdots) - f(0, 0, \cdots)| \le 2c_m \|f\|_\infty.$$

https://doi.org/10.1017/apr.2024.4 Published online by Cambridge University Press

Consequently, the event $(|X_i - X_i^{(m)}| \geq \epsilon)$ implies that $\epsilon \leq 2c_m\|f\|_\infty$, and then the probability $\mathbb{P}(|X_i - X_i^{(m)}| \geq \epsilon)$ vanishes whenever m satisfies $2c_m\|f\|_\infty < \epsilon$. We conclude that, for such m, $\mathbb{P}(d_H(\mathbb{X}_n^{(m)}, \mathbb{X}_n) \geq \epsilon) = 0$. □

We can now apply Proposition 4.1 to the m-dependent sequence $(X_n^{(m)})_{n\in\mathbb{Z}}$, combined with Lemma 4.1, to establish asymptotic (ϵ, α)-density for $(X_n)_{n\in\mathbb{Z}}$. Doing this, we obtain the result below under a suitable control of the following concentration coefficient, related to the truncated sequence $(X_n^{(m)})_{n\in\mathbb{Z}}$ (as defined in (4.3)):

$$\rho_{m+1}^{(m)}(\epsilon) = \inf_{x\in\mathbb{R}^{m+1}} \mathbb{P}(\|Y_{1,m+1}^{(m)} - x\| \leq \epsilon), \tag{4.5}$$

with $Y_{1,m+1}^{(m)} = (X_1^{(m)}, \cdots, X_{m+1}^{(m)})^t$.

Proposition 4.2. *Let $(X_n)_{n\in\mathbb{Z}}$ and f be as in the statement of Lemma 4.1. Let $\epsilon_0 > 0$, $\epsilon \in]0, \epsilon_0[$ be fixed, and let $m \in \mathbb{N}$ be such that $2c_m\|f\|_\infty < \epsilon$. Suppose that the concentration coefficient $\rho_{m+1}^{(m)}(\epsilon)$ related to the truncated sequence $(X_n^{(m)})_{n\in\mathbb{Z}}$, and defined in (4.5), satisfies*

$$\rho_{m+1}^{(m)}(\epsilon) \geq \kappa_\epsilon,$$

for some $\kappa_\epsilon > 0$. Then for any $n \geq m+1$,

$$\mathbb{P}(d_H(\mathbb{X}_n, \mathbb{M}) \geq 2\epsilon) \leq \frac{(1 - \kappa_{\frac{\epsilon}{2}})^{[\frac{1}{2}[\frac{n}{m+1}]]}}{\kappa_{\frac{\epsilon}{4}}},$$

and a conclusion similar to that of Proposition 4.1 is true for such m.

Proof. This follows from Lemma 4.1 together with Proposition 4.1 and the fact that

$$\mathbb{P}(d_H(\mathbb{X}_n, \mathbb{M}) \geq 2\epsilon) \leq \mathbb{P}(d_H(\mathbb{X}_n, \mathbb{X}_n^{(m)}) + d_H(\mathbb{X}_n^{(m)}, \mathbb{M}) \geq 2\epsilon)$$
$$\leq \mathbb{P}(d_H(\mathbb{X}_n, \mathbb{X}_n^{(m)}) \geq \epsilon) + \mathbb{P}(d_H(\mathbb{X}_n^{(m)}, \mathbb{M}) \geq \epsilon).$$

□

4.0.3. *Stationary β-mixing sequence on a compact set.* Recall that the stationary sequence $(X_n)_{n\in\mathbb{N}}$ is β-mixing if β_n tends to 0 when n tends to infinity, where the coefficients $(\beta_n)_{n>0}$ are defined as follows (see [5, 40] for the expression for β_n):

$$\beta_n = \sup_{l\geq 1} \mathbb{E}\left\{\sup |\mathbb{P}(B|\sigma(X_1, \cdots, X_l)) - \mathbb{P}(B)|\,,\ \ B \in \sigma(X_i,\ \ i \geq l+n)\right\}. \tag{4.6}$$

The following corollary gives conditions on the behavior of the two sequences $(\rho_n(\epsilon))_{n>0}$ and $(\beta_n)_{n>0}$ under which the asymptotically dense property of Definition 1.1 is satisfied.

Proposition 4.3. *Let $(X_n)_{n\geq 0}$ be a stationary β-mixing sequence. Suppose that X_1 is supported on a compact set $\mathbb{M}$. Then it holds, for any $\epsilon > 0$ and any sequences k_n and r_n such that $k_n r_n \leq n$,*

$$\mathbb{P}(d_H(\mathbb{X}_n, \mathbb{M}) > \epsilon) \leq \frac{k_n^2\beta_{r_n} + k_n \exp\left(-[\frac{k_n}{2}]\rho_{r_n}(\epsilon/2)\right)}{k_n\rho_{r_n}(\epsilon/4)}.$$

https://doi.org/10.1017/apr.2024.4 Published online by Cambridge University Press

Suppose moreover that for some $\beta > 1$, and any $\epsilon > 0$ small enough,

$$\lim_{m\to\infty} \rho_m(\epsilon)\frac{e^{m^\beta}}{m^{1+\beta}} = \infty, \quad \textit{and} \quad \lim_{m\to\infty} \frac{e^{2m^\beta}}{m^2}\beta_m = 0.$$

Then $(X_n)_{n\geq 0}$ is asymptotically dense in $\mathbb{M}$.

In the proof of Proposition 4.3 given in Section 7, we construct two sequences $(k_n)_n$ and $(r_n)_n$ for which

$$\lim_{n\to\infty} \frac{k_n^2\beta_{r_n} + k_n \exp\left(-[\frac{k_n}{2}]\rho_{r_n}(\epsilon/2)\right)}{k_n\rho_{r_n}(\epsilon/4)} = 0.$$

The threshold $n_0(\epsilon, \alpha)$ is not explicitly calculated, but it is that integer for which

$$\frac{k_n^2\beta_{r_n} + k_n \exp\left(-[\frac{k_n}{2}]\rho_{r_n}(\epsilon/2)\right)}{k_n\rho_{r_n}(\epsilon/4)} \leq \alpha,$$

for all $n \geq n_0(\epsilon, \alpha)$.

4.0.4. *Stationary weakly dependent sequence on a compact set.* We suppose here that $(X_i)_{i\in\mathbb{T}}$ is a stationary sequence such that X_1 takes values in a compact support $\mathbb{M}$. We suppose also that this sequence is weakly dependent in the sense of [14]. More precisely, we suppose that it satisfies the following definition.

Definition 4.1. We say that the sequence $(X_n)_{n\in\mathbb{T}}$ is $(\mathbb{L}_\infty, \Psi)$-weakly dependent if there exists a non-increasing function Ψ such that $\lim_{r\to\infty} \Psi(r) = 0$, and such that for any measurable functions f and g bounded (respectively by $\|f\|_\infty$ and $\|g\|_\infty$) and for any $i_1 \leq \cdots \leq i_k < i_k + r \leq i_{k+1} \leq \cdots \leq i_n$ one has

$$\left|\mathrm{Cov}\left(\frac{f(X_{i_1}, \cdots, X_{i_k})}{\|f\|_\infty}, \frac{g(X_{i_{k+1}}, \cdots, X_{i_n})}{\|g\|_\infty}\right)\right| \leq \Psi(r). \tag{4.7}$$

See [12, Definition 2.2] for a more general setting.

The dependence condition in Definition 4.1 is weaker than the Rosenblatt strong mixing dependence [35]. Let us briefly explain this. The α-mixing coefficient between the two sigma-fields $\mathcal{A}$ and $\mathcal{B}$ is defined as

$$\alpha(\mathcal{A}, \mathcal{B}) = \sup_{A\in\mathcal{A}, B\in\mathcal{B}} |\mathbb{P}(A \cap B) - \mathbb{P}(A)\mathbb{P}(B)|.$$

The sequence $(X_n)_{n\in\mathbb{T}}$ is strongly mixing if its coefficient α_n defined, for $n \geq 1$, by

$$\alpha_n = \sup_{k\in\mathbb{T}} \alpha(\mathcal{P}_k, \mathcal{F}_{k+n}),$$

tends to 0 as n tends to infinity, with $\mathcal{P}_k = \sigma(X_i,\ i \leq k)$ and $\mathcal{F}_{k+n} = \sigma(X_i,\ i \geq k+n)$. An equivalent formula for α_n, using the covariance between some functions, is stated in [6, Theorem 4.4]:

$$\alpha_n = \frac{1}{4}\sup\left\{\frac{\mathrm{Cov}(f, g)}{\|f\|_\infty\|g\|_\infty},\ f \in L_\infty(\mathcal{P}_k), g \in L_\infty(\mathcal{F}_{k+n})\right\},$$

https://doi.org/10.1017/apr.2024.4 Published online by Cambridge University Press

where $L_\infty(\mathcal{A})$ denotes the set of bounded $\mathcal{A}$-measurable functions for some σ-fields $\mathcal{A}$. It follows from this formula that strongly mixing sequences are $(\mathbb{L}_\infty, \Psi)$-weakly dependent, as stated in Definition 4.1 (with $\Psi(r) = \alpha_r$ for $r > 0$). The converse, however, is not necessarily true (see [12]).

We can now state our result for stationary weakly dependent sequences.

Proposition 4.4. *Let $(X_n)_{n\in\mathbb{T}}$ be a stationary, $(\mathbb{L}_\infty, \Psi)$-weakly dependent sequence in the sense of Definition 4.1. Suppose that X_1 is supported on a compact set $\mathbb{M}$. Then it holds, for any $\epsilon > 0$ and any sequences k_n and r_n such that $k_n r_n \le n$, that*

$$\mathbb{P}(d_H(\mathbb{X}_n, \mathbb{M}) > \epsilon) \le \frac{k_n^2 \Psi(r_n) + k_n \exp\left(-[\frac{k_n}{2}]\rho_{r_n}(\epsilon/2)\right)}{k_n \rho_{r_n}(\epsilon/4)}.$$

Suppose moreover that, for some $\beta > 1$, and any positive ϵ small enough,

$$\lim_{m\to\infty} \rho_m(\epsilon)\frac{e^{m^\beta}}{m^{1+\beta}} = \infty, \quad \text{and that} \quad \lim_{m\to\infty} \frac{e^{2m^\beta}}{m^2}\Psi(m) = 0.$$

Then this sequence $(X_n)_{n\in\mathbb{T}}$ is asymptotically dense in $\mathbb{M}$.

Here again the threshold $n_0(\epsilon, \alpha)$ is not explicitly calculated but can be given by an inequality, as in the case of β-mixing.

4.1. Comparison with the i.i.d. case

We can compare the bounds of Proposition 4.1 (Propositions 4.3 and 4.4 respectively) with what has already been obtained in the i.i.d. case (see [7, 10, 11]). That is, we restrict to the case when $m = 0$ (respectively, $k_n = n$ and $r_n = 1$, $\beta_n = \Psi(n) = 0$ for $n \ge 1$). Suppose that we are in this situation and that, moreover, $\rho_1(\epsilon)$ has a strictly positive lower bound, say κ_ϵ. Then all the conclusions of the three propositions above give the same upper bound for $\mathbb{P}(d_H(\mathbb{X}_n, \mathbb{M}) > \epsilon)$, which is

$$\frac{\exp(-[\frac{n}{2}]\kappa_{\frac{\epsilon}{2}})}{\kappa_{\frac{\epsilon}{4}}}. \tag{4.8}$$

Now we suppose, as already done in the i.i.d. case, that the (a, b)-standard assumption is satisfied, i.e, $\kappa_\epsilon = a\epsilon^b$, for some $a > 0$, $b > 0$, and for positive ϵ small enough. Recall that the (a, b)-standard assumption was used, in the i.i.d. context, for set estimation problems under Hausdorff distance ([10, 11]) and also for a statistical analysis of persistence diagrams ([7, 16]). Then, clearly, an upper bound for (4.8) is

$$C\frac{\exp(-cn\epsilon^b)}{\epsilon^b},$$

for some positive constants C and c (independent of n and of ϵ), as has already been found in the i.i.d. case (see for instance the upper bound (3.2) in [10]). Finally, we have to check that the requirements of Propositions 4.1, 4.3, and 4.4 are satisfied under the (a, b)-standard assumption (i.e. when $\rho_1(\epsilon) \ge a\epsilon^b$). Since we are in the case when $m = 0$ in Proposition 4.1, and when $\beta_n = 0$, $\Psi_n = 0$, $n \ge 1$ in Propositions 4.3 and 4.4, we have only to check that the (a, b)-standard assumption ensures, for i.i.d. random variables, that $\lim_{m\to\infty} \rho_m(\epsilon)\frac{e^{m^\beta}}{m^{1+\beta}} = \infty$. We deduce from the inequality

$$\|(X_1, \cdots, X_m)^t - (x_1, \cdots, x_m)^t\|^2 \le m \max_{1\le i\le m} \|X_i - x_i\|^2$$

https://doi.org/10.1017/apr.2024.4 Published online by Cambridge University Press

that

$$\mathbb{P}\left(m \max_{1\le i\le m} \|X_i - x_i\|^2 \le \epsilon^2\right) \le \mathbb{P}\left(\|(X_1, \cdots, X_m)^t - (x_1, \cdots, x_m)^t\|^2 \le \epsilon^2\right).$$

Now,

$$\mathbb{P}\left(m \max_{1\le i\le m} \|X_i - x_i\|^2 \le \epsilon^2\right) = \left(\mathbb{P}\left(\|X_1 - x_1\| \le \epsilon/\sqrt{m}\right)\right)^m.$$

Hence, $\rho_m(\epsilon) \ge \rho_1^m(\epsilon/\sqrt{m})$. Combining this bound with the (a, b)-standard assumption, we get

$$a^m \frac{\epsilon^{bm}}{m^{bm/2}} \frac{e^{m^\beta}}{m^{1+\beta}} \le \rho_m(\epsilon) \frac{e^{m^\beta}}{m^{1+\beta}}.$$

The left term tends to infinity as n goes to infinity (since $\beta > 1$); hence $\lim_{m\to\infty} \rho_m(\epsilon) \frac{e^{m^\beta}}{m^{1+\beta}} = \infty$.

As an important conclusion, the previous three propositions generalize the i.i.d. case well, even without the (a, b)-standard assumption.

5. Application to stationary Markov chains on a compact state space

This section gives conditions on stationary Markov chains on a compact state space that guarantee that they are asymptotically dense in this state space. Those conditions can be checked by studying the β-mixing properties of the Markov chains and applying Proposition 4.3 above. However, in this section we choose to be even more precise, adopting specific models and carrying out explicit calculations.

Let $(X_n)_{n\ge 0}$ be a homogeneous Markov chain satisfying the following two assumptions:

- $(\mathcal{A}_1)$ The Markov chain has an invariant measure μ with compact support $\mathbb{M}$ (and then the chain is stationary).
- $(\mathcal{A}_2)$ The transition probability kernel K, defined for $x \in \mathbb{M}$ by

  $$K(x, \cdot) = \mathbb{P}(X_1 \in \cdot | X_0 = x),$$

  is absolutely continuous with respect to some measure ν on $\mathbb{M}$; that is, there exist a positive measure ν and a positive function k such that for any $x \in \mathbb{M}$, $K(x, dy) = k(x, y)\nu(dy)$. Moreover, for some $b > 0$ and $\epsilon_0 > 0$,

  $$V_d := \inf_{x\in\mathbb{M}} \inf_{0<\epsilon<\epsilon_0} \left(\frac{1}{\epsilon^b} \int_{B(x,\epsilon)\cap\mathbb{M}} \nu(dx_1)\right) > 0$$

  and there exists a positive constant κ such that $\inf_{x\in\mathbb{M},\, y\in\mathbb{M}} k(x, y) \ge \kappa > 0$.

Proposition 5.1. *Suppose that Assumptions $(\mathcal{A}_1)$ and $(\mathcal{A}_2)$ are satisfied for some Markov chain $(X_n)_{n\ge 0}$. Then, for any $n \ge 1$ and any positive ϵ small enough,*

$$\mathbb{P}_\mu\left(d_H(\mathbb{X}_n, \mathbb{M}) > \epsilon\right) \le \frac{4^b(1 - \kappa\epsilon^b V_d/2^b)^n}{\kappa\epsilon^b V_d}.$$

https://doi.org/10.1017/apr.2024.4 Published online by Cambridge University Press

Consequently this Markov chain is asymptotically dense in $\mathbb{M}$*, with a threshold* $n_0(\epsilon, \alpha)$ *given by*

$$n_0(\epsilon, \alpha) = \frac{2^b}{\kappa \epsilon^b V_d} \left(\ln \left(\frac{4^b}{\kappa \epsilon^b V_d} \right) + \ln \left(\frac{1}{\alpha} \right) \right),$$

and V_d *is as introduced in Assumption* $(\mathcal{A}_2)$.

The proof and some key lemmas are deferred to Section 7.4.

We next give examples of Markov chains satisfying the requirements of Proposition 5.1. These examples concern stationary Markov chains on balls and stationary Markov chains on circles.

5.1. Stationary Markov chains on a ball of $\mathbb{R}^d$

5.1.1. *Random difference equations.* Let $(X_n)_{n\geq 0}$ be a Markov chain defined, for $n \geq 0$, by

$$X_{n+1} = A_{n+1} X_n + B_{n+1}, \tag{5.1}$$

where A_{n+1} is a $(d \times d)$-matrix, $X_n \in \mathbb{R}^d$, $B_n \in \mathbb{R}^d$, and $(A_n, B_n)_{n\geq 1}$ is an i.i.d. sequence independent of X_0. Recall that for a matrix M, $\|M\|$ is the operator norm, defined by $\|M\| = \sup_{x\in\mathbb{R}^d,\ \|x\|=1} \|Mx\|$. It is well known that for any $n \geq 1$, X_n is distributed as $\sum_{k=1}^{n} A_1 \cdots A_{k-1} B_k + A_1 \cdots A_n X_0$; see for instance [27]. It is also well known that the conditions (see [21, 28])

$$\mathbb{E}(\ln^+ \|A_1\|) < \infty, \quad \mathbb{E}(\ln^+ \|B_1\|) < \infty, \quad \lim_{n\to\infty} \frac{1}{n} \ln \|A_1 \cdots A_n\| < 0 \ \text{ a.s.} \tag{5.2}$$

ensure the existence of a stationary solution to (5.1), and that $\|A_1 \cdots A_n\|$ approaches 0 exponentially fast. If in addition $\mathbb{E}\|B_1\|^\beta < \infty$ for some $\beta > 0$, then the series $R := \sum_{i=1}^{\infty} A_1 \cdots A_{i-1} B_i$ converges a.s. and the distribution of X_n converges to that of R, independently of X_0. The distribution of R is then that of the stationary measure of the chain.

Compact state space. If $\|B_1\| \leq c < \infty$ for some fixed c, then this stationary Markov chain is $\mathbb{M}$-compactly supported. In particular, if $\|A_1\| \leq \rho < 1$ for some fixed ρ, then $\mathbb{M}$ is included in the ball $B_d(0, \frac{c}{1-\rho})$ of $\mathbb{R}^d$.

Transition kernel. Suppose that, for any $x \in \mathbb{M}$, the random vector $A_1 x + B_1$ has a density $f_{A_1 x + B_1}$ with respect to the Lebesgue measure (here ν is the Lebesgue measure) satisfying $\inf_{x,\ y\in\mathbb{M}} f_{A_1 x + B_1}(y) \geq \kappa$; then $k(x, y) = f_{A_1 x + B_1}(y) \geq \kappa > 0$.

We collect all of the above results in the following corollary.

Corollary 5.1. *Suppose that in the model* (5.1)*, the conditions* (5.2) *are satisfied, and moreover* $\|B_1\| \leq c < \infty$*. If the density of* $A_1 x + B_1$*, denoted by* $f_{A_1 x + B_1}$*, satisfies* $\inf_{x,\ y\in\mathbb{M}} f_{A_1 x + B_1}(y) \geq \kappa > 0$ *for some positive* κ*, then Assumptions* $(\mathcal{A}_1)$ *and* $(\mathcal{A}_2)$ *are satisfied with* $b = d$ *and* ν *the Lebesgue measure on* $\mathbb{R}^d$*.*

Example: the AR(1) process in $\mathbb{R}$. We consider a particular case of the Markov chain as defined in (5.1) with $d = 1$, where, for each n, $A_n = \rho$ with $|\rho| < 1$. We obtain the standard first-order linear autoregressive process, i.e.,

$$X_{n+1} = \rho X_n + B_{n+1}.$$

https://doi.org/10.1017/apr.2024.4 Published online by Cambridge University Press

We suppose that

- B_1 has a density function f_B supported on $[-c, c]$ for some $c > 0$ with $\kappa := \inf_{x\in[-c,c]} f_B(x) > 0$, and
- $X_0 \in [\frac{-c}{1-|\rho|}, \frac{c}{1-|\rho|}]$.

This Markov chain evolves in a compact state space which is a subset of $[\frac{-c}{1-|\rho|}, \frac{c}{1-|\rho|}]$. Thanks to Corollary 5.1, $(X_n)_n$ admits a stationary measure μ. We have, moreover,

$$k(x, y) = f_{B_1}(y - \rho x) \geq \kappa, \quad \forall\, x \in \mathbb{M}, \quad \forall\, y \in \mathbb{M}.$$

Assumptions $(\mathcal{A}_1)$ and $(\mathcal{A}_2)$ are then satisfied with $b = 1$ and ν the Lebesgue measure on $\mathbb{R}$.

Example: the AR(k) process in $\mathbb{R}$. The AR(k) process is defined by

$$Y_n = \alpha_1 Y_{n-1} + \alpha_2 Y_{n-2} + \cdots + \alpha_k Y_{n-k} + \epsilon_n,$$

where $\alpha_1, \cdots, \alpha_k \in \mathbb{R}$. Since this model can be written in the form of (5.1) with $d = 1$,

$$X_n = (Y_n, Y_{n-1}, \cdots, Y_{n-k+1})^t, \quad B_n = (\epsilon_n, 0, \cdots, 0)^t, \quad A_n = \begin{pmatrix} \alpha_1 & \cdots & \alpha_k \\ I_{k-1} & & 0 \end{pmatrix},$$

all of the above results for random difference equations apply under the corresponding assumptions. In particular, the process AR(2) is stationary as soon as $|\alpha_2| < 1$ and $\alpha_2 + |\alpha_1| < 1$.

5.2. The Möbius Markov chain on the circle

Our aim is to give an example of a Markov chain on the unit circle, known as the Möbius Markov chain, which satisfies the requirements of Proposition 5.1. This Markov chain $(X_n)_{n\in\mathbb{N}}$ is introduced in [26] and is defined as follows:

- X_0 is a random variable which takes values on the unit circle.
- For $n \geq 1$,

$$X_n = \frac{X_{n-1} + \beta}{\beta X_{n-1} + 1}\epsilon_n,$$

where $\beta \in]-1, 1[$ and $(\epsilon_n)_{n\geq 1}$ is a sequence of i.i.d. random variables which are independent of X_0 and distributed as the wrapped Cauchy distribution with a common density with respect to the arc length measure ν on the unit circle $\partial B(0, 1)$,

$$f_\varphi(z) = \frac{1}{2\pi}\frac{1-\varphi^2}{|z-\varphi|^2}, \quad \varphi \in [0, 1[\text{ fixed}, \quad z \in \partial B(0, 1).$$

The following proposition holds.

Proposition 5.2. *Let $(X_n)_{n\geq 0}$ be the Möbius Markov chain on the unit circle as defined above. Then this Markov chain admits a unique invariant distribution, denoted by μ. If X_0 is distributed as μ, then the set $\mathbb{X}_n = \{X_1, \cdots, X_n\}$ converges in probability, as n tends to infinity,*

https://doi.org/10.1017/apr.2024.4 Published online by Cambridge University Press

in the Hausdorff distance to the unit circle $\partial B(0,1)$*; more precisely, for any* $\alpha \in]0,1[$*, any positive* ϵ *sufficiently small, and any* $n \geq \frac{2}{\kappa v \epsilon}\left(\ln(\frac{1}{\alpha}) + \ln(\frac{4}{\epsilon \kappa v})\right)$*, we have*

$$d_H(\mathbb{X}_n, \partial B(0,1)) \leq \epsilon$$

with probability at least $1-\alpha$*. Here* v *is a finite positive constant and* $\kappa = \frac{1}{2\pi}\frac{1-\varphi}{1+\varphi}$.

The Möbius Markov chain of Proposition 5.2 is then asymptotically dense in the unit circle with a threshold $n_0(\epsilon, \alpha)$ given by

$$n_0(\epsilon, \alpha) = \frac{2}{\kappa v \epsilon}\left(\ln\left(\frac{1}{\alpha}\right) + \ln\left(\frac{4}{\epsilon \kappa v}\right)\right),$$

κ being as in the statement of the proposition, while the positive constant v is defined by (5.5) below.

Proof. We have to prove that all the requirements of Proposition 5.1 are satisfied. Our main reference for this proof is [26], where it is shown that this Markov chain has a unique invariant measure μ on the unit circle. This measure μ has full support on $\partial B(0,1)$ (so that Assumption $(\mathcal{A}_1)$ is satisfied with $\mathbb{M} = \partial B(0,1)$). The task now is to check Assumption $(\mathcal{A}_2)$. We have also, for $x \in \partial B(0,1)$,

$$K(x, dz) = \mathbb{P}(X_1 \in dz | X_0 = x) = k(x,z)\nu(dz), \tag{5.3}$$

where ν is the arc length measure on the unit circle, and for $x, z \in \partial B(0,1)$,

$$k(x,z) = \frac{1}{2\pi}\frac{1-|\phi_1(x)|^2}{|z-\phi_1(x)|^2},$$

with

$$\phi_1(x) = \frac{\varphi x + \beta\varphi}{\beta x + 1}.$$

Since $\frac{x+\beta}{\beta x+1} \in \partial B(0,1)$ whenever $x \in \partial B(0,1)$, we obtain $|\phi_1(x)|^2 = \varphi^2$. Now, for $x, z \in \partial B(0,1)$,

$$|z - \phi_1(x)| \leq |z| + |\phi_1(x)| \leq 1 + \varphi.$$

Hence,

$$k(x,z) \geq \frac{1}{2\pi}\frac{1-\varphi^2}{(1+\varphi)^2} = \frac{1}{2\pi}\frac{1-\varphi}{1+\varphi} > 0. \tag{5.4}$$

We now have to check that, for some $\epsilon_0 > 0$,

$$v := \inf_{u \in \partial B(0,1)} \inf_{0<\epsilon<\epsilon_0} \left(\epsilon^{-1} \int_{\partial B(0,1) \cap B(u,\epsilon)} \nu(dx_1)\right) > 0. \tag{5.5}$$

For this let $u \in \partial B(0,1)$, and define $\widehat{AB} = \int_{\partial B(0,1)\cap B(u,\epsilon)} \nu(dx_1)$.

https://doi.org/10.1017/apr.2024.4 Published online by Cambridge University Press

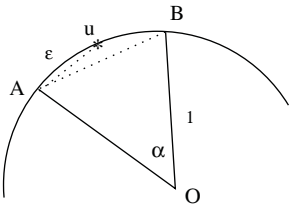

FIGURE 3. The ball $B(u, \epsilon)$ intersects the unit circle at two points A and B.

We have $\|u - A\| = \|u - B\| = \epsilon$ (see Figure 3). Let $\alpha = \widehat{AOB}$; then on the one hand $\overset{\frown}{AB} = \alpha$. On the other hand, since the triangle OAu is isosceles, with an angle of $\alpha/2$ in O, we have $\epsilon = 2\sin(\alpha/4)$. We thus obtain

$$\lim_{\epsilon \to 0} \frac{1}{\epsilon} \overset{\frown}{AB} = \lim_{\epsilon \to 0} \frac{\alpha}{\epsilon} = \lim_{\alpha \to 0} \frac{\alpha}{2\sin(\alpha/4)} = 2.$$

From this, (5.5) is satisfied.

Assumption $(\mathcal{A}_2)$ is satisfied thanks to (5.3), (5.4), and (5.5). The proof of Proposition 5.2 is then complete by Proposition 5.1. □

6. Simulations

The purpose of this section is to simulate a Möbius Markov process on the unit circle (as defined in Section 5.2) and to illustrate the results of Proposition 5.2 and Theorem 2.2. More precisely, we simulate the following:

- a random variable, X_0, distributed as the uniform law on the unit circle $\partial B(0, 1)$; that is, X_0 has the density

$$f(z) = \frac{1}{2\pi}, \quad \forall\ z \in \partial B(0, 1);$$

- for $n \geq 1$, $X_n = X_{n-1}\epsilon_n,$, where $(\epsilon_n)_{n \geq 1}$ is a sequence of i.i.d. random variables which are independent of X_0 and distributed as the wrapped Cauchy distribution with a common density, with respect to the arc length measure ν on the unit circle $\partial B(0, 1)$, given by

$$f_\varphi(z) = \frac{1}{2\pi} \frac{1 - \varphi^2}{|z - \varphi|^2}, \quad \varphi \in [0, 1[, \quad z \in \partial B(0, 1).$$

In this case, it is proved in [26] that this Markov chain is stationary and its stationary measure is the uniform law on the unit circle.

https://doi.org/10.1017/apr.2024.4 Published online by Cambridge University Press

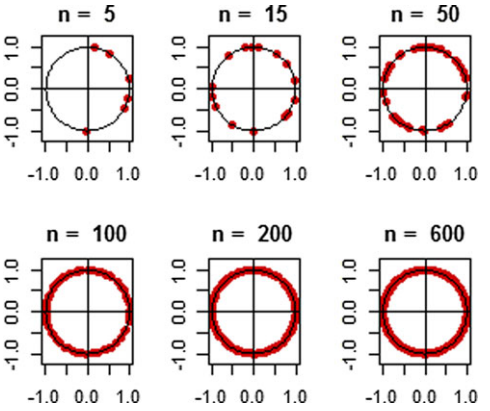

FIGURE 4. Illustrations of the set $\{x_1, \cdots, x_n\}$ which is a realization of the stationary random variables $\mathbb{X}_n = \{X_1, \ldots, X_n\}$ for different values of n and with $\varphi = 0$.

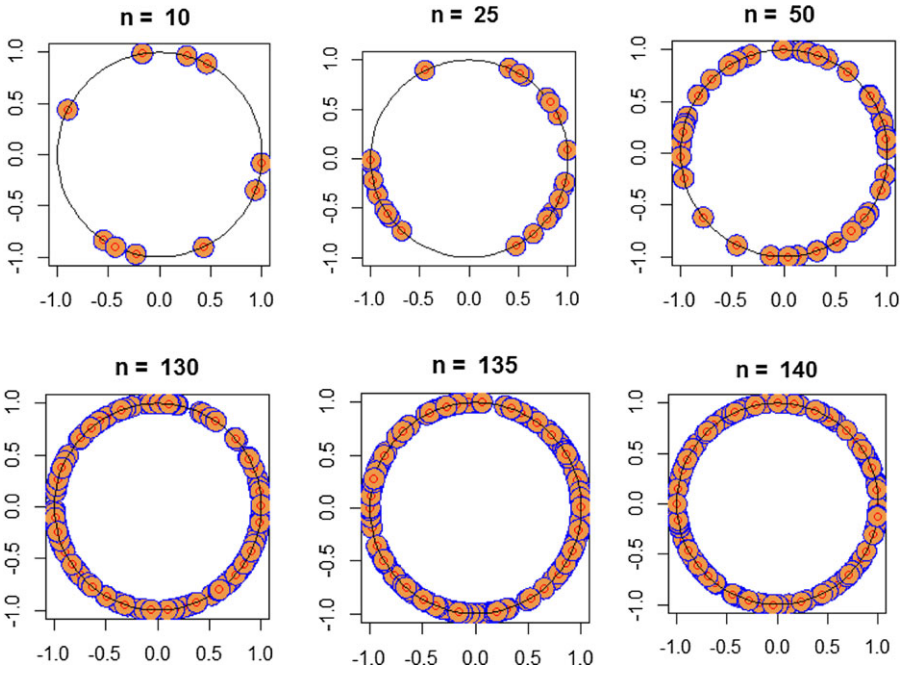

FIGURE 5. In the above images, the points of $\mathbb{X}_n$ are in red. Each of these points is the center of a circle with radius $r = 0.1$. This is an illustration of the reconstruction result $M \xrightarrow{\simeq} \bigcup_{x \in \mathbb{X}_n} B(x, r)$, with different values of n and with $r = 0.1$. In the above, there is reconstruction when $n = 140$ and $r = 0.1$. The density $\frac{\epsilon}{2}$ is at least $\frac{2\pi}{280} = 0.0224$, and so ϵ is at least 0.0448. For this value of $\epsilon = 0.0448$, $\epsilon < r = 0.1$ and this reconstruction is consistent with Theorem 2.2.

https://doi.org/10.1017/apr.2024.4 Published online by Cambridge University Press

7. Deferred proofs

7.1. Proof of Proposition 4.1

Let $\epsilon_0 > 0$ and $\epsilon \in]0, \epsilon_0[$ be fixed. In this proof we set $m' = m + 1$, $r = m'$, and $k = k_n = [n/m']$. Proposition 3.1, applied with these values of r and k, gives

$$\mathbb{P}(d_H(\mathbb{X}_n, \mathbb{M}) > \epsilon) \leq \mathbb{P}\Big(d_H(\{Y_{1,m'}, \cdots, Y_{k_n,m'}\}, \mathbb{M}_{dm'}) > \epsilon\Big)$$

$$\leq \frac{\sup_{x \in \mathbb{M}_{dm'}} \mathbb{P}\Big(\min_{1 \leq i \leq k_n} \|Y_{i,m'} - x\| > \epsilon/2\Big)}{1 - \sup_{x \in \mathbb{M}_{dm'}} \mathbb{P}\Big(\|Y_{1,m'} - x\| > \epsilon/4\Big)}, \tag{7.1}$$

where $Y_{i,m'} = (X_{(i-1)m'+1}, \cdots, X_{im'})^t$. The sequence $(X_n)_{n \in \mathbb{T}}$ is stationary and assumed to be m-dependent. Consequently, the two families $\{Y_{1,m'}, Y_{3,m'}, Y_{5,m'}, \cdots\}$ and $\{Y_{2,m'}, Y_{4,m'}, Y_{6,m'}, \cdots\}$ each consist of i.i.d. random vectors. Since we are assuming that $\rho_{m'}(\epsilon) \geq \kappa_\epsilon$, we have

$$\sup_{x \in \mathbb{M}_{dm'}} \mathbb{P}\left(\min_{1 \leq i \leq k_n} \|Y_{i,m'} - x\| > \frac{\epsilon}{2}\right) \leq \sup_{x \in \mathbb{M}_{dm'}} \mathbb{P}\left(\min_{1 \leq 2i \leq k_n} \|Y_{2i,m'} - x\| > \frac{\epsilon}{2}\right)$$

$$\leq \sup_{x \in \mathbb{M}_{dm'}} \left(\mathbb{P}\left(\|Y_{1,m'} - x\| > \frac{\epsilon}{2}\right)\right)^{[k_n/2]} \leq \left(1 - \rho_{m'}(\frac{\epsilon}{2})\right)^{[k_n/2]} \leq \left(1 - \kappa_{\frac{\epsilon}{2}}\right)^{[k_n/2]} \tag{7.2}$$

and

$$1 - \sup_{x \in \mathbb{M}_{dr}} \mathbb{P}\Big(\|Y_{1,m'} - x\| > \frac{\epsilon}{4}\Big) \geq \kappa_{\frac{\epsilon}{4}}. \tag{7.3}$$

Collecting the bounds (7.1), (7.2), and (7.3), we find that for any $\epsilon > 0$,

$$\mathbb{P}(d_H(\mathbb{X}_n, \mathbb{M}) > \epsilon) \leq \frac{\left(1 - \kappa_{\frac{\epsilon}{2}}\right)^{[k_n/2]}}{\kappa_{\frac{\epsilon}{4}}} \leq \frac{\exp\left(-\kappa_{\frac{\epsilon}{2}}[k_n/2]\right)}{\kappa_{\frac{\epsilon}{4}}}.$$

Let $\alpha \in]0, 1[$ be such that $\dfrac{\exp\left(-\kappa_{\frac{\epsilon}{2}}[k_n/2]\right)}{\kappa_{\frac{\epsilon}{4}}} \leq \alpha$, which is equivalent to

$$[k_n/2] \geq \frac{1}{\kappa_{\frac{\epsilon}{2}}} \log\left(\frac{1}{\alpha \kappa_{\frac{\epsilon}{4}}}\right).$$

Then, for any $n \geq \frac{2m'}{\kappa_{\frac{\epsilon}{2}}} \log\left(\frac{1}{\alpha \kappa_{\frac{\epsilon}{4}}}\right) + 3m'$, we have

$$[k_n/2] \geq k_n/2 - 1 \geq \frac{n}{2m'} - 3/2 \geq \frac{1}{\kappa_{\frac{\epsilon}{2}}} \log\left(\frac{1}{\alpha \kappa_{\frac{\epsilon}{4}}}\right),$$

and therefore $\mathbb{P}(d_H(\mathbb{X}_n, \mathbb{M}) > \epsilon) \leq \alpha$. The proof of Proposition 4.1 is complete. □

https://doi.org/10.1017/apr.2024.4 Published online by Cambridge University Press

7.2. Proof of Proposition 4.3

We use the blocking method of [40] to transform the dependent β-mixing sequence $(X_n)_{n\in\mathbb{N}}$ into a sequence of nearly independent blocks. Let $Z_{2i,r_n} = (\xi_j,\ j \in \{(2i-1)r_n + 1, \cdots, 2ir_n\})^t$ be a sequence of i.i.d. random vectors independent of the sequence $(X_i)_{i\in\mathbb{N}}$ such that, for any i, Z_{2i,r_n} is distributed as Y_{2i,r_n} (which is distributed as Y_{1,r_n}). Lemma 4.1 of [40] proves that the two vectors $(Z_{2i,r_n})_i$ and $(Y_{2i,r_n})_i$ are related by the following relation:

$$\left|\mathbb{E}(h(Z_{2i,r_n},\, 1 \le 2i \le k_n)) - \mathbb{E}(h(Y_{2i,r_n},\, 1 \le 2i \le k_n))\right| \le k_n\beta_{r_n},$$

which is true for any measurable function bounded by 1. We then have, using the last bound,

$$\begin{aligned}
&k_n \sup_{x\in\mathbb{M}_{dr_n}} \mathbb{P}\left(\min_{1\le i\le k_n} \|Y_{i,r_n} - x\| > \epsilon\right) \le k_n \sup_{x\in\mathbb{M}_{dr_n}} \mathbb{P}\left(\min_{1\le 2i\le k_n} \|Y_{2i,r_n} - x\| > \epsilon\right)\\
&\le k_n \sup_{x\in\mathbb{M}_{dr_n}} \left|\mathbb{P}\left(\min_{1\le 2i\le k_n} \|Y_{2i,r_n} - x\| > \epsilon\right) - \mathbb{P}\left(\min_{1\le 2i\le k_n} \|Z_{2i,r_n} - x\| > \epsilon\right)\right|\\
&+k_n \sup_{x\in\mathbb{M}_{dr_n}} \mathbb{P}\left(\min_{1\le 2i\le k_n} \|Z_{2i,r_n} - x\| > \epsilon\right)\\
&\le k_n^2\beta_{r_n} + k_n \sup_{x\in\mathbb{M}_{dr_n}} \mathbb{P}\left(\min_{1\le 2i\le k_n} \|Z_{2i,r_n} - x\| > \epsilon\right)\\
&\le k_n^2\beta_{r_n} + k_n \sup_{x\in\mathbb{M}_{dr_n}} \left(\mathbb{P}\big(\|Y_{1,r_n} - x\| > \epsilon\big)\right)^{[k_n/2]}\\
&\le k_n^2\beta_{r_n} + k_n\left(1 - \rho_{r_n}(\epsilon)\right)^{[k_n/2]}\\
&\le k_n^2\beta_{r_n} + k_n \exp\left(-[\frac{k_n}{2}]\rho_{r_n}(\epsilon)\right)
\end{aligned}$$

and

$$1 - \sup_{x\in\mathbb{M}_{dr_n}} \mathbb{P}\big(\|Y_{1,r_n} - x\| > \epsilon/4\big) = \rho_{r_n}(\epsilon/4).$$

Consequently, Proposition 3.1 gives

$$\begin{aligned}
\mathbb{P}(d_H(\mathbb{X}_n, \mathbb{M}) > \epsilon) &\le \frac{\sup_{x\in\mathbb{M}_{dr_n}} \mathbb{P}\big(\min_{1\le i\le k_n} \|Y_{i,r_n} - x\| > \epsilon/2\big)}{1 - \sup_{x\in\mathbb{M}_{dr_n}} \mathbb{P}\big(\|Y_{1,r_n} - x\| > \epsilon/4\big)}\\
&\le \frac{k_n^2\beta_{r_n} + k_n \exp\left(-[\frac{k_n}{2}]\rho_{r_n}(\epsilon/2)\right)}{k_n\rho_{r_n}(\epsilon/4)}.
\end{aligned} \tag{7.4}$$

We now have to construct two sequences k_n and r_n such that $k_n r_n \le n$ and

$$\lim_{n\to\infty} k_n^2\beta_{r_n} = 0,\quad \lim_{n\to\infty} k_n\rho_{r_n}(\epsilon) = \infty,\quad \lim_{n\to\infty} k_n \exp\left(-\frac{k_n}{2}\rho_{r_n}(\epsilon)\right) = 0. \tag{7.5}$$

https://doi.org/10.1017/apr.2024.4 Published online by Cambridge University Press

We have supposed that $\lim_{m\to\infty} \rho_m(\epsilon)\frac{e^{m^\beta}}{m^{1+\beta}} = \infty$ for some $\beta > 1$. Define $\gamma = 1/\beta \in]0, 1[$ and

$$k_n = \left[\frac{n}{(\ln n)^\gamma}\right], \quad r_n = [(\ln n)^\gamma].$$

Then, letting $m = r_n = [(\ln n)^\gamma]$, we have $\lim_{n\to\infty} k_n \frac{\rho_{r_n}(\epsilon)}{\ln n} = \infty$, and then (since $k_n \le n$),

$$\lim_{n\to\infty} k_n \frac{\rho_{r_n}(\epsilon)}{\ln (k_n)} = \infty.$$

From the last limit we have that $\lim_{n\to\infty} k_n \rho_{r_n}(\epsilon) = \infty$ and that, for n large enough and for some $C > 2$, $k_n \frac{\rho_{r_n}(\epsilon)}{\ln (k_n)} \ge C$, so that

$$k_n \exp\left(-\frac{k_n}{2}\rho_{r_n}(\epsilon)\right) \le k_n^{1-C/2}.$$

Consequently, $\lim_{n\to\infty} k_n \exp\left(-\frac{k_n}{2}\rho_{r_n}(\epsilon)\right) = 0$. Now we deduce from $\lim_{m\to\infty} \frac{e^{2m^\beta}}{m^2}\beta_m = 0$ that (letting $m = r_n = [(\ln n)^\gamma]$)

$$\lim_{n\to\infty} k_n^2 \beta_{r_n} = 0.$$

The two sequences k_n and r_n so constructed satisfy (7.5), and for these sequences it holds that

$$\lim_{n\to\infty} \frac{k_n^2\beta_{r_n} + k_n \exp\left(-\frac{k_n}{2}\rho_{r_n}(\epsilon/2)\right)}{k_n \rho_{r_n}(\epsilon/4)} = 0.$$

Hence for any $\alpha \in]0, 1[$ there exists an integer $n_0(\epsilon, \alpha)$ such that for any $n \ge n_0(\epsilon, \alpha)$,

$$\frac{k_n^2\beta_{r_n} + k_n \exp\left(-\frac{k_n}{2}\rho_{r_n}(\epsilon/2)\right)}{k_n \rho_{r_n}(\epsilon/4)} \le \alpha.$$

Combining this last inequality with that of (7.4) finishes the proof of Proposition 4.3. □

7.3. Proof of Proposition 4.4

We have

$$\begin{aligned} k_n \mathbb{P}\left(\min_{1\le i\le k_n} \|Y_{i,r_n} - x\| > \epsilon\right) &\le k_n \mathbb{P}\left(\min_{1\le 2i\le k_n} \|Y_{2i,r_n} - x\| > \epsilon\right) \\ &\le k_n \left| \mathbb{P}\left(\min_{1\le 2i\le k_n} \|Y_{2i,r_n} - x\| > \epsilon\right) - \prod_{i:\, 1\le 2i\le k_n} \mathbb{P}\left(\|Y_{2i,r_n} - x\| > \epsilon\right) \right| \\ &+ k_n \prod_{i:\, 1\le 2i\le k_n} \mathbb{P}\left(\|Y_{2i,r_n} - x\| > \epsilon\right). \end{aligned} \tag{7.6}$$

https://doi.org/10.1017/apr.2024.4 Published online by Cambridge University Press

For s events $A_1, \cdots, A_s$ (with the convention that, $\prod_{j=1}^{0} \mathbb{P}(A_j) = 1$), we have

$$\mathbb{P}(A_1 \cap \cdots \cap A_s) - \prod_{i=1}^{s} \mathbb{P}(A_i) = \sum_{i=1}^{s-1} \mathbb{P}(A_1) \cdots \mathbb{P}(A_{i-1}) \mathrm{Cov}(\mathbb{1}_{A_i}, \mathbb{1}_{A_{i+1} \cap \cdots \cap A_s}).$$

Hence,

$$\left| \mathbb{P}(A_1 \cap \cdots \cap A_s) - \prod_{i=1}^{s} \mathbb{P}(A_i) \right| \leq \sum_{i=1}^{s-1} \left| \mathrm{Cov}(\mathbb{1}_{A_i}, \mathbb{1}_{A_{i+1} \cap \cdots \cap A_s}) \right|.$$

Applying the last bound with $A_i = \left(\|Y_{2i,r_n} - x\| > \epsilon \right)$ and using (4.7), we get

$$\left| \mathrm{Cov}(\mathbb{1}_{A_i}, \mathbb{1}_{A_{i+1} \cap \cdots \cap A_s}) \right| \leq \Psi(r_n)$$

and

$$\left| \mathbb{P}\left(\min_{1 \leq 2i \leq k_n} \|Y_{2i,r_n} - x\| > \epsilon \right) - \prod_{i:\, 1 \leq 2i \leq k_n} \mathbb{P}\left(\|Y_{2i,r_n} - x\| > \epsilon \right) \right| \leq k_n \Psi(r_n). \tag{7.7}$$

Combining (7.6) and (7.7), we deduce that

$$k_n \mathbb{P}\left(\min_{1 \leq i \leq k_n} \|Y_{i,r_n} - x\| > \epsilon \right) \leq k_n^2 \Psi(r_n) + k_n \left(1 - \rho_{r_n}(\epsilon) \right)^{[k_n/2]}$$

$$\leq k_n^2 \Psi(r_n) + k_n \exp\left(-[k_n/2] \rho_{r_n}(\epsilon) \right).$$

Consequently, as for (7.4), we get

$$\mathbb{P}(d_H(\mathbb{X}_n, \mathbb{M}) > \epsilon) \leq \frac{k_n^2 \Psi(r_n) + k_n \exp\left(-[\frac{k_n}{2}] \rho_{r_n}(\epsilon/2) \right)}{k_n \rho_{r_n}(\epsilon/4)}.$$

We now have to construct two sequences r_n and k_n such that

$$\lim_{n \to \infty} k_n \exp(-k_n \rho_{r_n}(\epsilon)/2) = 0, \quad \lim_{n \to \infty} k_n^2 \Psi(r_n) = 0, \quad \lim_{n \to \infty} k_n \rho_{r_n}(\epsilon) = \infty.$$

This last construction is possible as argued at the end of the proof of Proposition 4.3. □

7.4. Lemmas for Section 5

To prove Proposition 5.1, we need the following two lemmas in order to check the conditions of Proposition 3.1 (with $r = 1$). Recall that $\mathbb{P}_x$ (resp. $\mathbb{P}_\mu$) denotes the probability when the initial condition $X_0 = x$ (resp. X_0 is distributed as the stationary measure μ).

Lemma 7.1. *Let $(X_n)_{n \geq 0}$ be a Markov chain satisfying Assumptions $(\mathcal{A}_1)$ and $(\mathcal{A}_2)$. Then, for any $0 < \epsilon < \epsilon_0$ and any $x_0 \in \mathbb{M}$, it holds that*

$$\inf_{x \in \mathbb{M}} \mathbb{P}_{x_0}\left(\|X_1 - x\| \leq \epsilon \right) \geq \kappa \epsilon^b V_d, \quad \inf_{x \in \mathbb{M}} \mathbb{P}_\mu\left(\|X_1 - x\| \leq \epsilon \right) \geq \kappa \epsilon^b V_d.$$

https://doi.org/10.1017/apr.2024.4 Published online by Cambridge University Press

Proof. Using Assumption $(\mathcal{A}_2)$, we have

$$\mathbb{P}_{x_0}\left(\|X_1 - x\| \le \epsilon\right) = \mathbb{P}_{x_0}\left(X_1 \in B(x,\epsilon) \cap \mathbb{M}\right) = \int_{B(x,\epsilon)\cap\mathbb{M}} K(x_0, dx_1)$$

$$= \int_{B(x,\epsilon)\cap\mathbb{M}} k(x_0, x_1)\nu(dx_1)$$

$$\ge \kappa \int_{B(x,\epsilon)\cap\mathbb{M}} \nu(dx_1) \ge \kappa\epsilon^b \inf_{0<\epsilon<\epsilon_0}\left(\frac{1}{\epsilon^b}\int_{B(x,\epsilon)\cap\mathbb{M}} \nu(dx_1)\right) \ge \kappa\epsilon^b V_d.$$

The proof of Lemma 7.1 is then complete since $\mathbb{P}_\mu\left(\|X_1 - x\| \le \epsilon\right) = \int \mathbb{P}_{x_0}\left(\|X_1 - x\| \le \epsilon\right) d\mu(x_0)$.

Lemma 7.2. *Let $(X_n)_{n\ge 0}$ be a Markov chain satisfying Assumptions $(\mathcal{A}_1)$ and $(\mathcal{A}_2)$. Then, for any $0 < \epsilon < \epsilon_0$ and $k \in \mathbb{N}\setminus\{0\}$, it holds that*

$$\sup_{x\in\mathbb{M}} \mathbb{P}_\mu\left(\min_{1\le i\le k}\|X_i - x\| > \epsilon\right) \le \left(1 - \kappa\epsilon^b V_d\right)^k.$$

Proof. Set $\mathcal{F}_n = \sigma(X_0, \ldots, X_n)$. By the Markov property and Lemma 7.1,

$$\mathbb{P}_\mu\left(\min_{1\le i\le k}\|X_i - x\| > \epsilon\right) = \mathbb{P}_\mu(\forall\, 1 \le i \le k,\ X_i \notin B(x,\epsilon))$$

$$= \mathbb{E}_\mu\left(\prod_{i=1}^{k-1} \mathbb{1}_{\{X_i\notin B(x,\epsilon)\}}\mathbb{E}(\mathbb{1}_{\{X_k\notin B(x,\epsilon)\}}|\mathcal{F}_{k-1})\right)$$

$$= \mathbb{E}_\mu\left(\prod_{i=1}^{k-1} \mathbb{1}_{\{X_i\notin B(x,\epsilon)\}}\mathbb{E}_{X_{k-1}}(\mathbb{1}_{\{X_k\notin B(x,\epsilon)\}})\right)$$

$$\le (1 - \kappa\epsilon^b V_d)\mathbb{E}_\mu\left(\prod_{i=1}^{k-1} \mathbb{1}_{\{X_i\notin B(x,\epsilon)\}}\right)$$

$$\le (1 - \kappa\epsilon^b V_d)\mathbb{P}_\mu(\forall\, 1 \le i \le k-1,\ X_i \notin B(x,\epsilon)).$$

Lemma 7.2 is proved using the last bound together with an argument by induction on k. □

7.5. Proof of Proposition 5.1

Proposition 3.1 *, applied with $r = r_n = 1$ and $k = k_n = n$, gives*

$$\mathbb{P}_\mu\left(d_H(\mathbb{X}_n, \mathbb{M}) > \epsilon\right) \le \frac{\sup_{x\in\mathbb{M}_d}\mathbb{P}_\mu\left(\min_{1\le i\le n}\|Y_{i,1} - x\| > \epsilon/2\right)}{1 - \sup_{x\in\mathbb{M}_d}\mathbb{P}_\mu\left(\|Y_{1,1} - x\| > \epsilon/4\right)},$$

with $Y_{i,r} = (X_{(i-1)r+1}, \cdots, X_{ir})$ so that $Y_{i,1} = X_i$. Consequently, noting that $\mathbb{M}_d = \mathbb{M}$, we have

$$\mathbb{P}_\mu\left(d_H(\mathbb{X}_n, \mathbb{M}) > \epsilon\right) \le \frac{\sup_{x\in\mathbb{M}}\mathbb{P}_\mu\left(\min_{1\le i\le n}\|X_i - x\| > \epsilon/2\right)}{1 - \sup_{x\in\mathbb{M}}\mathbb{P}_\mu\left(\|X_1 - x\| > \epsilon/4\right)}.$$

https://doi.org/10.1017/apr.2024.4 Published online by Cambridge University Press

Now Lemmas 7.1 and 7.2 give

$$\sup_{x\in\mathbf{M}} \mathbb{P}_\mu \left(\min_{1\le i\le n} \|X_i - x\| > \epsilon \right) \le (1 - \kappa\epsilon^b V_d)^n \le \exp(-n\kappa\epsilon^b V_d),$$

$$1 - \sup_{x\in\mathbf{M}} \mathbb{P}_\mu \left(\|X_1 - x\| > \epsilon \right) \ge \kappa\epsilon^b V_d > 0.$$

Combining the three last bounds, we obtain

$$\mathbb{P}_\mu \left(d_H(\mathbb{X}_n, \mathbf{M}) > \epsilon \right) \le \frac{4^b \exp(-n\kappa\epsilon^b V_d/2^b)}{\kappa\epsilon^b V_d}.$$

The proof of the proposition is then complete since $\alpha \ge \dfrac{4^b \exp(-n\kappa\epsilon^b V_d/2^b)}{\kappa\epsilon^b V_d}$ *is equivalent to*

$$n \ge \frac{2^b}{\kappa\epsilon^b V_d} \ln\left(\frac{4^b}{\alpha\kappa\epsilon^b V_d} \right).$$

Acknowledgements

We are very grateful to both referees for their insightful comments, and for suggesting many improvements. One of the referees made very thorough and relevant remarks on both the topological and probabilistic parts. The first author would like to thank Sebastian Scholtes for insightful discussions on the material of Section 2 and [36]. The second author is grateful to Sophie Lemaire for the present form of the proof of Lemma 7.2.

Funding information

The first author is supported by an SFRG grant from the American University of Sharjah (AUS, UAE). The second author is supported by the Université Grenoble Alpes, CNRS, LJK.

Competing interests

There were no competing interests to declare which arose during the preparation or publication process of this article.

References

[1] Aamari, E. and Levrard, C. (2019). Nonasymptotic rates for manifold, tangent space and curvature estimation, *Ann. Statist.* **47**, 177–204.
[2] Amézquita, E. J. *et al.* (2020). The shape of things to come: topological data analysis and biology, from molecules to organisms. *Dev. Dynamics* **249**, 816–833.
[3] Attali, D., Lieutier, A. and Salinas, D. (2013). Vietoris–Rips complexes also provide topologically correct reconstructions of sampled shapes. *Comput. Geom.* **46**, 448–465.
[4] Barb, S. (2009). Topics in geometric analysis with applications to partial differential equations. Doctoral Thesis, University of Missouri-Columbia.
[5] Bradley, R. C. (1983). Absolute regularity and functions of Markov chains. *Stoch. Process. Appl.* **14**, 67–77.
[6] Bradley, R. C. (2007). *Introduction to Strong Mixing Conditions*, Vol. 1, 2, 3. Kendrick Press, Heber City, UT.
[7] Chazal, F., Glisse, M., Labruyère, C. and Michel, B. (2015). Convergence rates for persistence diagram estimation in topological data analysis. *J. Machine Learning Res.* **16**, 3603–3635.
[8] Chazal, F. and Oudot, S. Y. (2008). Towards persistence-based reconstruction in Euclidean spaces. In *Scg '08: Proceedings of the Twenty-Fourth Annual Symposium on Computational Geometry*, Association for Computing Machinery, New York, pp. 232–241.

https://doi.org/10.1017/apr.2024.4 Published online by Cambridge University Press

[9] CISEWSKI-KEHE, J. *et al.* (2018). Investigating the cosmic web with topological data analysis. In *American Astronomical Society Meeting Abstracts* **231**, id. 213.07.
[10] CUEVAS, A. (2009). Set estimation: another bridge between statistics and geometry. *Bol. Estadist. Investig. Oper.* **25**, 71–85.
[11] CUEVAS, A. AND RODRGUEZ-CASAL, A. (2004). On boundary estimation. *Adv. Appl. Prob.* **36**, 340–354.
[12] DEDECKER, J. *et al.* (2007). *Weak Dependence: With Examples and Applications*. Springer, New York.
[13] DIVOL, V. (2021). Minimax adaptive estimation in manifold inference. *Electron. J. Statist.* **15**, 5888–5932.
[14] DOUKHAN, P. AND LOUHICHI, S. (1999). A new weak dependence condition and applications to moment inequalities. *Stoch. Process. Appl.* **84**, 313–342.
[15] ELLIS, J. C. (2012). On the geometry of sets of positive reach. Doctoral Thesis, University of Georgia.
[16] FASY, B. T. *et al.* (2014). Confidence sets for persistence diagrams. *Ann. Statist.* **42**, 2301–2339.
[17] FEDERER, H. (1959). Curvature measures. *Trans. Amer. Math. Soc.* **93**, 418–491.
[18] FEFFERMAN, C., MITTER, S. AND NARAYANAN, H. (2016). Testing the manifold hypothesis. *J. Amer. Math. Soc.* **29**, 983–1049.
[19] FU, J. H. G. (1989). Curvature measures and generalized Morse theory. *J. Differential Geom.* **30**, 619–642.
[20] INIESTA, R. *et al.* (2022). Topological Data Analysis and its usefulness for precision medicine studies. *Statist. Operat. Res. Trans.* **46**, 115–136.
[21] GOLDIE, C. M. AND MALLER, R. A. (2001). Stability of perpetuities. *Ann. Prob.* **28**, 1195–1218.
[22] HOEF, L. V., ADAMS, H., KING, E. J. AND EBERT-UPHOFF, I. (2023). A primer on topological data analysis to support image analysis tasks in environmental science. *Artificial Intellig. Earth Systems* **2**, 1–18.
[23] HATCHER, A. (2002.) *Algebraic Topology*. Cambridge University Press.
[24] HÖRMANDER, L. (1994). *Notions of Convexity*. Birkhäuser, Boston.
[25] HÖRMANN, S. AND KOKOSZKA, P. (2010). Weakly dependent functional data. *Ann. Statist.* **38**, 1845–1884.
[26] KATO S. (2010). A Markov process for circular data. *J. R. Statist. Soc. B. [Statist. Methodology]* **72**, 655–672.
[27] KESTEN, H. (1973). Random difference equations and renewal theory for products of random matrices. *Acta Math.* **131**, 207–248.
[28] KESTEN, H. (1974). Renewal theory for functionals of a Markov chain with general state space. *Ann. Prob.* **2**, 355–386.
[29] KIM, J. *et al.* (2020). Homotopy reconstruction via the Cech complex and the Vietoris–Rips complex. In *36th International Symposium on Computational Geometry*, Dagstuhl Publishing, Wadern, article no. 54.
[30] LEE, J. M. (2013). *Introduction to Smooth Manifolds*. Springer, New York.
[31] MATHOVERFLOW (2017). Tubular neighborhood theorem for C^1 submanifold. Available at https://mathoverflow.net/questions/286512/tubular-neighborhood-theorem-for-c1-submanifold.
[32] MORENO, R., KOPPAL, S. AND DE MUINCK, E. (2013). Robust estimation of distance between sets of points. *Pattern Recognition Lett.* **34**, 2192–2198.
[33] NIYOGI, P., SMALE, S. AND WEINBERGER, S. (2008). Finding the homology of submanifolds with high confidence from random samples. *Discrete Comput. Geom.* **39**, 419–441.
[34] RIO, E. (2013). Inequalities and limit theorems for weakly dependent sequences. Available at https://cel.hal.science/cel-00867106v2.
[35] ROSENBLATT, M. (1956). A central limit theorem and a strong mixing condition. *Proc. Nat. Acad. Sci. USA* **42**, 43–47.
[36] SCHOLTES, S. (2013). On hypersurfaces of positive reach, alternating Steiner formulae and Hadwiger's Problem. Preprint. Available at https://arxiv.org/abs/1304.4179.
[37] SINGH, Y. *et al.* (2023). Topological data analysis in medical imaging: current state of the art. *Insights Imaging* **14**, article no. 58.
[38] THALE, C. (2008). 50 years sets with positive reach—a survey. *Surveys Math. Appl.* **3**, 123–165.
[39] WANG, Y. AND WANG, B. (2020). Topological inference of manifolds with boundary. *Comput. Geom.* **88**, article no. 101606.
[40] YU, B. (1994). Rates of convergence for empirical processes of stationary mixing sequences. *Ann. Prob.* **22**, 94–116.

https://doi.org/10.1017/apr.2024.4 Published online by Cambridge University Press